Stages en Deep Learning pour la médecine personnalisée : Prédictions, Interprétation de modèle, Modèles génératifs, Apprentissage par transfert

La structure que vous allez rejoindre

Le laboratoire IBISC (Informatique, Bioinformatique et Systèmes Complexes) de l'université Paris-Saclay (Univ. Evry) est une structure de recherche comprenant une centaine de chercheurs de haut niveau. Vous rejoindrez une équipe d'une dizaine **d'experts en machine learning**. Notre principal champ de recherche concerne les modèles de **deep learning** et en particulier les modèles de prédiction, l'apprentissage par transfert, l'interprétation de modèle, l'intégration de connaissances, les modèles génératifs. Une grande partie des recherches effectuées sont appliquées à l'analyse de **données génomiques et à la médecine personnalisée**. Nous publions nos résultats de recherche dans des articles scientifiques publiés dans les meilleurs journaux et conférences de machine learning et bioinformatique. Nous collaborons avec de nombreuses structures académiques (IRD, INSERM, Sorbonne université, …) et entreprises (SANOFI, IRT SYSTEMX, start-up, …).

Contexte

L'apprentissage profond **(Deep Learning)** est une avancée majeure de l'intelligence artificielle de ces dernières années. Cette approche de l'apprentissage automatique consiste à apprendre à un réseau de neurones de grande taille à réaliser une tâche de prédiction à l'aide d'un ensemble de données d'apprentissage. L'apprentissage profond s'est rapidement imposé comme un standard dans plusieurs domaines en pulvérisant les records des précédentes méthodes de l'état de l'art. Ses domaines de prédilection sont principalement l'analyse d'images et le traitement du langage naturel. Un des futurs enjeux majeurs de cette approche est son **application à la médecine personnalisée**.

Nos thèmes de recherche se concentrent plus spécifiquement sur la **prédiction de phénotypes** (diagnostiques, pronostiques, réponse aux traitements…) à partir de **données d'expression génique** (transcriptomiques). L'expression génique représente l'activité des gènes d'un individu quantifié par le nombre de brins d'ARN identifié dans un prélèvement. Les modèles de réseaux de neurones profonds permettent de prendre en compte la grande complexité et les nombreuses interactions entre gènes pour calculer des prédictions fiables. Notre équipe souhaite explorer des nouvelles voies de recherche à travers **trois stages** concernant : les modèles génératifs, l'interprétation de modèles et l'apprentissage à partir de peu de données.

Sujet 1 : Modèle de diffusion pour la génération de données d'expression

Les **modèles de diffusion** sont des modèles génératifs récents qui ont été remarqués par leur utilisation dans les modèles « text to image » tels que **Dalle2** ou **Midjourney**. Ces méthodes semblent surpasser les performances de l'état de l'art constitué des réseaux adversariaux génératifs (GAN) et des autoencodeurs variationels (VAE). Leur principe est d'ajouter itérativement à une image un faible bruit, puis d'utiliser toutes ces images bruitées pour apprendre un modèle qui enlève le bruit d'une image. En appliquant ce modèle de nombreuses fois sur des données aléatoires, on finit par générer une image de grande qualité.

L'objectif de ce stage est d'adapter les modèles de diffusion à la génération de données d'expression de gènes. Les enjeux de concevoir un générateur de données transcriptomiques sont triples : 1) Les jeux de données sont généralement de petite taille car chers à produire, un générateur permettrait de faire de **l'augmentation de données** et donc d'améliorer les performances des modèles de prédiction. 2) En fournissant des données artificielles grâce au générateur, les chercheurs pourront développer des modèles sans avoir besoin d'accéder aux données réelles et donc préserver la **confidentialité** des données patients. 3) A l'aide d'un générateur, il sera possible de **simuler des interventions** sur l'expression des gènes et d'identifier de potentielles cibles thérapeutiques pour certaines maladies.

[1] Goodfellow, Ian, et al. "Generative adversarial networks." (2014)

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851.

[3] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

[4] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

[5] Rombach, Robin et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022): 10674-10685.

Sujet 2 : Interprétation contrefactuelle de modèle profond pour la découverte de biomarqueurs

Le deep learning est appelé à jouer un rôle majeur dans le diagnostic ou la décision thérapeutique. Les modèles appris à partir de données génomiques permettent de prédire divers phénotypes de patients. Étant donné que ces modèles sont très précis, il serait pertinent de comprendre sur quels éléments la décision du modèle est basée. Les variables les plus significatives pour ces modèles pourraient être considérées comme des potentiels **biomarqueurs** ou cibles thérapeutiques de la maladie.

L'idée de ce sujet de stage est qu'une **interprétation contrefactuelle** d'un réseau de neurones pourrait identifier des biomarqueurs pertinents. La première étape sera de construire un modèle qui prédit avec une grande précision un phénotype donné (type de cancer, pronostic). Ensuite, nous chercherons des explications contrefactuelles aux prédictions fournies par le modèle. Cela se fera par une procédure d'optimisation où nous recherchons la transformation minimale d'une donnée patient qui modifie la prédiction du modèle. Plusieurs contraintes doivent être ajoutées ; par exemple, le patient contrefactuel doit être dans la distribution réelle des données du patient, ou les corrélations entre les variables doivent être préservées. Ce problème d'optimisation devra être résolu par **apprentissage des exemples contrefactuels** en gelant les paramètres du modèle profond. Des biomarqueurs potentiels seront identifiés en analysant la différence entre les patients réels et contrefactuels.

[1] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Journal*, 2017.

[2] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-Objective Counterfactual Explanations," 2020.

[3] S. Verma, J. Dickerson, and K. Hines, "Counterfactual Explanations for Machine Learning: A Review." arXiv, 2020.

[4] A. Van Looveren and J. Klaise, "Interpretable Counterfactual Explanations Guided by Prototypes." arXiv, 2020.

• Sujet 3 : Apprentissage à partir de peu de données

Un des verrous scientifiques majeurs à lever pour avancer dans ce domaine est l'apprentissage de réseaux de neurones à partir de **jeux d'apprentissage de petite taille**. En effet, dû à leur coût élevé d'acquisition, les jeux de données transcriptomiques publics contiennent très peu de patients étiquetés (quelques milliers au mieux). À cause de ce faible nombre d'exemples, l'apprentissage des réseaux de neurones profonds se heurte à des problèmes de sur-apprentissage, le réseau apprend par cœur les données mais pas le concept sous-jacent.

L'objectif de ce sujet est de pallier le problème des jeux de données de petite taille en ayant recours à des approches basées sur l'**apprentissage par transfert** [1] (adaptation de domaine ou auto-supervision). Le principe général est d'apprendre une représentation optimale des données dans les couches cachées d'un réseau qui sera ensuite utilisée pour rendre la tâche de prédiction initiale plus facile. Spécifiquement, l'approche d'**adaptation de domaine** vise à apprendre des modèles en intégrant plusieurs jeux de données issus de distributions différentes, par exemple des données de patients et de lignées cellulaires. En s'inspirant des méthodes telles que CyCADA [2] ou DANN [3], nous pourrions aligner les différentes sources de données pour apprendre un modèle prédictif commun, ou appliquer le « style » d'une source A vers une source B et ainsi agrandir la base d'apprentissage. Dans le cadre de l'approche d'**auto-supervision**, le modèle est pré-appris à partir d'un large jeu de données non étiqueté en travaillant sur une tâche dite prétexte. Une partie du modèle ainsi pré-appris sera ajustée sur un jeu de données cible proche du jeu de données non étiqueté. Nous pourrions explorer les méthodes récentes développées sur les données tabulaires telles que VIME [4] ou SubTab [5].

[1] Hanczar, B., Bourgeais, V., & Zehraoui, F. (2022). Assessment of deep learning and transfer learning for cancer prediction based on gene expression data. *BMC Bioinformatics*, *23*(1), 262.

[2] Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., ... & Darrell, T. (2018, July). CyCADA: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning* (pp. 1989-1998). Pmlr.

[3] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096-2030.

[4]. Yoon, J., Zhang, Y., Jordon, J., & van der Schaar, M. (2020). VIME : Extending the Success of Self- and Semi-supervised Learning to Tabular Domain. *Advances in Neural Information Processing Systems*, *33*, 11033-11043.

[5] Ucar, T., Hajiramezanali, E., & Edwards, L. (2021). SubTab : Subsetting Features of Tabular Data for Self-Supervised Representation Learning. *Advances in Neural Information Processing Systems*, *34*, 18853-18865.

Déroulement du stage :

- Lecture et synthèse de l'état de l'art scientifique du domaine
- Sélection et adaptation des méthodes les plus pertinentes pour l'application en génomique
- Programmation de la solution proposée en Python / Pytorch
- Test et évaluation sur des jeux de données publiques et éventuellement privée en fonction de l'application
- Rédaction d'un article scientifique si les résultats sont concluants

Le profil que nous recherchons :

- Master 2 ou dernière année d'école d'ingénieur en informatique ou mathématiques appliquées.
- De solides compétences en apprentissage profond sur une diversité d'architectures et de problématiques appliqués aux données (signaux, images, texte, etc.)
- Connaissances des librairies de machine learning (Scikit-learn) et de deep learning (Pytorch)
- Si possible un goût et une expérience sur les problématiques méthodologiques liés à la santé ou génomique.
- Motivé par la recherche scientifique
- Travail en équipe, communication et créativité

Pour proposer votre candidature

- Type de contrat : Stage 5-6 mois
- Poursuite en thèse possible
- Lieux : Evry (télétravail partiel possible)
- Début du contrat : 2023
- Envoyer votre CV et les relevés de notes de vos deux derrières années à blaise.hanczar[at]univ-evry.fr