Sujets de stage recherche Master 2 / Ecole d'Ingénieur

Apprentissage profond appliqué aux données transcriptomiques : « transfer learning, self-supervised, domain adaptation »

Laboratoire IBISC – Université Paris-Saclay (Evry)

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 santé.

Nos thèmes de recherche se concentrent plus spécifiquement sur la prédiction de phénotypes (diagnostiques, pronostiques, réponse aux traitements…) à partir de données d'expression de gènes. Un verrou scientifique majeur à lever pour avancer dans ce domaine est l'apprentissage de réseaux de neurones à partir de jeux d'apprentissage de petite taille. Nous proposons des stages liés à ce sujet à partir d'approches auto-supervisées (*self-supervised learning*) et d'adaptation de domaine.

Sujet :

L'analyse de données transcriptomiques par apprentissage profond est un domaine de recherche très récent. La grande majorité des articles publiés a moins de deux ans et parmi eux seulement une poignée s'intéresse à la prédiction de phénotypes. La raison de ce faible nombre de travaux publiés actuellement provient du manque de grands jeux de données transcriptomiques disponibles dû à leur coût élevé d'acquisition. Alors que les réseaux de neurones profonds traitant des images ou du langage naturel sont construits à partir de plusieurs centaines de milliers ou millions d'exemples, les jeux de données transcriptomiques publiques contiennent très peu de patients (quelques milliers au mieux). A 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.

Nous comptons pallier le problème de la petite taille des données d'apprentissage en utilisant différentes approches. Dans les approches *self-supervised*, nous ferons le pré-apprentissage d'un modèle à partir d'un large jeu de données non étiqueté. Des méthodes récentes de contrastive learning [1], teacher-student [2] ou de clustering self-supervised [3] seront adaptées au problème des données transcriptomiques. Les approches d'adaptation de domaines, nous permettrons d'apprendre des modèles avec plusieurs jeux de données issues de distributions différentes, par exemples des données de patients et de lignées cellulaires. En s'inspirant de méthodes tel que CADA [4] ou DANN [5], nous devrons aligner les différentes sources de données pour apprendre un modèle prédictif commun. Dans les deux approches, le principe est d'apprendre une représentation optimale des données dans les couches cachées du réseau et qui sera utilisée afin de rendre la tâche de prédiction initiale plus facile.

Dans ces stages, nous utiliserons ces méthodes pour transférer de l'information à travers plusieurs réseaux appris à partir de petits jeux de données transcriptomiques dans le but d'améliorer les performances de prédictions. Le travail consistera à faire un état de l'art et à sélectionner les méthodes d'apprentissage les plus performantes actuellement sur les données images. Puis il faudra adapter les approches sélectionnées pour une utilisation sur les données transcriptomiques. La dernière étape sera de tester les méthodes développées à travers une série d'expérimentations sur des jeux de données publiques.

Cadre :

Ces stages se dérouleront au laboratoire IBISC de l'université Paris-Saclay (13ème université mondiale du classement Shangaï 2021) dans l'équipe AROBAS. Les recherches de l'équipe AROBAS se concentrent entre autres sur l'apprentissage automatique et la bio-informatique. Depuis 2015, un de ses thèmes majeurs est le développement de nouvelles méthodes d'apprentissage profond pour l'analyse de données génomiques.

Profils recherchés :

- Etudiant Master 2 Recherche ou en dernière année d'école d'ingénieur de formation informatique ou mathématiques appliquées.
- Une solide formation en machine learning est indispensable.
- Des bases en programmation python et une bonne maîtrise de l'anglais sont nécessaires.
- Des connaissances en deep learning et programmation tensorflow / pytorch seraient appréciées.
- Autonomie et curiosité pour la recherche scientifique.

Début du stage : 2022
Durée : 5-6 mois

Encadrant : Pr Hanczar Blaise

Pour postuler envoyer CV et relevé de notes à blaise.hanczar@ibisc.univ-evry.fr

Références :

[1] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big Self-Supervised Models are Strong Semi-Supervised Learners. Advances in Neural Information Processing Systems, 33.

[2] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap Your Own Latent—A New Approach to Self-Supervised Learning. Advances in Neural Information Processing Systems, 33, 21271–21284.

[3] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 9912–9924). Curran Associates, Inc.

[4] Zou, H., Zhou, Y., Yang, J., Liu, H., Das, H. P., & Spanos, C. J. (2019). Consensus Adversarial Domain Adaptation. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 5997–6004.

[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2017). Domain-Adversarial Training of Neural Networks. In G. Csurka (Ed.), Domain Adaptation in Computer Vision Applications (pp. 189–209). Springer International Publishing.