Multimodal Pre-Training for Scientific Data

François Lanusse

24/05/2024

Abstract

L'apprentissage profond a récemment connu un changement de paradigme, passant de l'entraînement de modèles spécialisés sur des jeux de données dédiés, aux modèles dits de Fondation, entraînés de manière auto-supervisée sur de vastes quantités de données, puis adaptés pour résoudre des tâches spécifiques avec des performances optimales. Ce nouveau paradigme a eu un très grand succès pour les grands modèles de langage (LLMs) mais aussi dans d'autres domaines tels que les modèles de vision. Cependant, les applications de cette nouvelle approche dans les sciences physiques sont encore très rares, pour des raisons allant de la nouveauté des techniques associées, à l'absence de grands ensembles de données appropriés à l'entraînement de modèles de fondation.

Dans cette présentation, je discuterai de nos travaux récents sur le déploiement d'une telle approche dans le contexte de l'astrophysique observationnelle. Notre objectif est d'intégrer une grande quantité d'observations hétérogènes (par exemple, différents types de mesures, différents instruments, etc.) dans un même système d'IA, capable d'isoler les propriétés physiques des objets observés, avec un minimum de supervision.

Je présenterai dans un premier temps les résultats de notre première génération de modèles (que nous appelons AstroCLIP) qui nous permet de créer de manière non supervisée une représentation jointe de données d'imagerie et de spectroscopie astronomique. Cette représentation peut ensuite être utilisée pour diverses applications en aval (par exemple, l'estimation du décalage vers le rouge, la classification morphologique, l'estimation des propriétés physiques) avec des méthodes d'apprentissage automatique très simples tout en atteignant des performances optimales.

Je discuterai ensuite des différents étapes que nous entreprenons pour le développement d'une seconde génération de modèles à plus grande échelle, incluant la création du plus grand jeux de données astronomique pour l'IA disponible à ce jour, le développement de techniques nous permettant de prendre en compte les metadatas des données scientifiques et d'ingérer dans un même système des données variés (images, séries temporelles, spectres, données tabulaires), et l'utilisation de modèles de transformers dits multi-modaux capables de modéliser les corrélations entres ces différents types d'observations.