Rapport final de projet

Étude de modèle d'intelligence artificielle en sismologie

Groupe 1

Julien LABORDE-PEYRÉ, Enio LAHITTE, Yuhan MA, Timothée COLETTE

CentraleSupélec

LMPS Laboratoire de Mécanique Paris-Saclay

Janvier 2025

Sommaire

1 Introduction

Ce rapport final a pour but de recueillir l'ensemble des éléments qui ont constitué notre projet. Celui-ci contient, tout d'abord, une présentation de notre client et du problème posé, un état de l'art complet et précis sur la situation, l'ensemble des tâches réalisées dans le cadre de notre projet, pour conclure sur une réflexion et une prise de recul concernant les possibles perspectives de ce dernier.

1.1 Présentation du client

Notre client est M. Filippo GATTI, enseignant-chercheur à CentraleSupélec au sein du laboratoire de mécanique de Paris-Saclay (LMPS). Le LMPS comprend plus de 220 membres, dont de nombreux doctorants, ingénieurs, post-doctorants et émérites. Au sein du LMPS, M. GATTI réalise plusieurs travaux, notamment des recherches en ingénierie sismique et, ce qui nous intéressera davantage ici, l'application de l'intelligence artificielle à la propagation des ondes. Filippo GATTI supervise notre projet et est notre interlocuteur privilégié lors de nos réunions. Il nous aide à définir les besoins et nous oriente sur différentes pistes à étudier en nous suggérant des algorithmes à tester ou des bases de données intéressantes pour fine-tuner nos modèles.

1.2 Identification du problème posé

Notre client nous a proposé d'étudier deux modèles de fondation déjà existant, le premier : "Seismic Foundation Model" prend en entrée des images et réalise des tâches diverses telles que

- **Classification** : Classe les différentes couches de terrain dans les images sismiques.
- **Segmentation** : Sépare et identifie les formes géologiques spécifiques (comme des poches de pétrole ou de gaz).
- **Inversion** : Analyse les données pour comprendre les types de roches et leurs propriétés.
- **Signal prossessing** : Nettoie les données pour enlever les interférences et les rendre plus lisibles.
- **Interpolation** : Remplit les trous dans les données pour obtenir une image complète.

Le deuxième : SeisLM, prend quant à lui en entré des signaux simsiques il réalise des tâches telles que :

- **Détection d'événéments** : Détection de signaux d'événements ou de bruit.
- **Identification de phase** : Identification d'onde P ou S.
- **Prédiction de Onset** : Prévision des valeurs futures d'une série temporelle.

Actuellement, les deux modèles ne prennent en entrée qu'un seul format de données. Ajouter d'autres types de données pourrait améliorer la qualité des prédictions. Par exemple, combiner des relevés sismiques avec des images pour des tâches de segmentation d'imagerie sismique serait pertinent, car la vitesse de propagation des ondes dépend des matériaux traversés. Ainsi, l'un des objectifs de notre projet est de fusionner ces deux modèles pour créer un modèle multimodal.
La plus-value attendue par notre client sur le projet réside, tout d'abord, dans la mesure de

l'efficacité du modèle sur les tâches présentées précédemment, ainsi que dans la création d'un outil de visualisation simple permettant un accès rapide aux réponses du modèle. Ensuite, elle repose sur l'implémentation d'un modèle multimodal qui prend, dans notre cas, en entrée à la fois des signaux et des images. Une fois ce modèle implémenté, nous devrons également évaluer son efficacité afin de le comparer à l'ancien modèle et déterminer si la combinaison des deux modalités améliore les performances ou, au contraire, engendre de la confusion.

2 État de l'art

L'objectif de ce projet est de développer un modèle multimodal performant en s'appuyant sur deux modèles de fondation existants, [**SFM**][1] et [**SeisLM**][2]. Ces modèles ont démontré leur efficacité respective dans le traitement des signaux sismiques pour [**SeisLM**][2],et le traitement des topographies en 3 dimensions pour [**SFM**][1]. Cependant, chacun présente des limites spécifiques, notamment le fait qu'ils n'acceptent qu'un type de donnée d'entrée. Notre approche vise à combiner leurs forces complémentaires pour créer un nouveau modèle capable de traiter des données de nature différente, en améliorant la robustesse et la précision des prédictions. En intégrant les capacités des deux modèles et en les adaptant à notre problématique, nous espérons apporter un petit plus dans le domaine des modèles de fondation multimodaux.

Premièrement, notre projet repose en grande partie sur les travaux de l'équipe [**SFM**][1]. Ce projet se veut expérimental, car il figure parmi les premiers d'une telle envergure dans le domaine de la géophysique. Après avoir présenté les différentes fonctionnalités, l'architecture et les performances de leur modèle, ces travaux proposent plusieurs perspectives d'évolution dans une section "Discussion". C'est notamment dans cette partie que l'idée de développer un modèle multimodal est évoquée : "Multimodal Geoscience Models : Training language models to guide image models is an emerging trend. By combining language with geoscience data, we can train models more efficiently." Cette mention a particulièrement inspiré notre encadrant, M. Gatti, à nous proposer ce projet.La publication [**SFM**][1] a notamment précédé celle de [**SeisLM**][2], publiée très récemment, le 21 octobre 2024. Il est intéressant de noter que le second article cite le premier dans ses références. Alors que [**SFM**][1] se concentre principalement sur l'exploitation des données de topographie en 3D, [**SeisLM**][2] explore l'utilisation de signaux temporels tridimensionnels issus de relevés expérimentaux d'appareils de sismologie, lesquels représentent les secousses dans les trois directions de l'espace en fonction du temps. Cet article illustre qu'il est possible d'extraire directement des informations pertinentes à partir de ces signaux, bien que les tâches traitées diffèrent. En effet, [**SeisLM**][2] se focalise sur la prédiction d'événements sismiques et la classification des ondes P et S, dans le cadre du picking d'ondes. Toutefois, cette démonstration renforce l'idée que ces signaux sismiques peuvent être utilisés pour générer des connaissances exploitables. Par ailleurs, ces mêmes signaux constituent une base essentielle pour les ensembles de données utilisés par [**SFM**][1], car la reconstitution 3D des sols repose sur l'envoi d'ondes sismiques dans le sol et l'interprétation de leur comportement.

D'autres modèles de fondations multimodaux ont déjà vu le jour, on peut citer par exemple [**PROSE-FD**][4], cependant ce modèle s'inscrit dans un tout autre domaine, celui de la mécanique des fluides. Le modèle prend en entrée d'une part les conditions initiales physiques des milieux (modalité 1) et d'autre part les équations aux dérivées partielles qui caractérisent l'évolution du système (modalité 2). Cet article peut être une piste inspirante pour voir comment combiner dans un même modèle deux types de données complètement différentes puisqu'en effet leur module intégré "fusion" en sortie pourrait correspondre à ce qu'il faudrait que l'on utilise. Mais cet article montre également que, de manière plus générale, l'utilisation d'entrées multimodales peut s'avérer cruciale. Dans ce cas précis, les équations

fournissent le "sens physique", c'est-à-dire l'évolution du système, tandis que les données initiales déterminent le point de départ de la simulation. Ces deux blocs de données sont interdépendants, chacun étant indispensable à l'autre pour assurer une modélisation complète et précise. Cela met en évidence que la multimodalité peut être une composante essentielle dans le cadre d'un projet, en particulier lorsque différentes sources d'information se complètent pour fournir une vision plus riche et pertinente du système étudié.

D'un point de vue plus technique, une de nos pistes d'exploration fut l'utilisation d'un unique modèle de fondation, le modèle SFM dans notre cas, et de l'entraîner avec de nouvelles données provenant d'une modalité différente, à savoir les signaux temporels. Pour ce faire, il est nécessaire de prétraiter les signaux afin de les convertir en données compatibles avec le modèle. Nous avons ainsi décidé de transformer les signaux en images de 224x224 pixels (compatibles avec le modèle) à l'aide de la transformation en spectrogramme de Mel. Cette technique a été utilisée dans l'article [**MEL in ASR**][5] dans le cadre du développement d'un transformeur pour la reconnaissance automatique de la parole (ASR), permettant de convertir des enregistrements audio en images. Ce modèle a obtenu des résultats de précision supérieurs à ceux des autres modèles d'ASR, démontrant l'efficacité de l'ajout d'une modalité sous forme de signaux temporels convertis en images via le spectrogramme de Mel. Ainsi, cet exemple d'utilisation du spectrogramme de Mel nous encourage à appliquer cette stratégie à notre modèle de fondation SFM.

Sur un plan plus général, l'article [**Multimodal Foundation Models**][3] traite très largement des modèles de foundation multimodaux. Il souligne que ce domaine connaît une transition significative, permettant de passer de modèles spécialisés à des assistants généralistes capables de traiter des tâches variées. L'article offre une vue d'ensemble de cette évolution, en identifiant deux grandes catégories : les modèles pré-entraînés pour des tâches spécifiques et les assistants généraux. Ces derniers, inspirés par les avancées des grands modèles linguistiques (LLMs), cherchent à unifier les capacités de compréhension et de génération visuelles dans une architecture cohérente. Cet article semble être un peu trop technique et ambitieux dans le cadre de notre projet, cependant il montre que l'implémentation de modèle de fondation multimodaux est quelque chose de pertinent et d'intéressant.

Pour mieux comprendre l'utilité et le fonctionnement des transformeurs dans notre projet, nous nous sommes appuyés sur l'article [**An image is worth 16x16 words**][6] publié en 2021 par des chercheurs de Google, ainsi que sur les travaux liés au modèle SFM dans le domaine de la géophysique.

Dans cet article, les auteurs démontrent que les transformeurs, bien connus pour leur performance en traitement du langage naturel, peuvent également exceller en vision par ordinateur. En divisant une image en petits morceaux appelés patches, le modèle applique un mécanisme d'attention pour analyser les relations entre ces fragments. Ce processus permet de mettre en évidence les zones les plus pertinentes de l'image, rendant le modèle particulièrement performant pour des tâches telles que la classification ou la reconnaissance d'images.

Grâce à cet article, nous avons pris conscience du potentiel des transformeurs pour la reconnaissance d'images en utilisant des patches. En particulier, le modèle ViT (Vision Trans-

former) nous a semblé pertinent, car il permet de mieux comprendre les relations globales entre les différentes parties d'une image, contrairement aux réseaux de neurones convolutifs (CNN) qui effectuent une analyse principalement locale. De plus, grâce à son mécanisme d'attention, ViT est capable de capturer des structures complexes dans les données visuelles. Il se révèle donc essentiel pour des tâches nécessitant une analyse spatiale approfondie, comme la classification et la reconnaissance d'images. Notons également que ViT montre des performances particulièrement impressionnantes lorsqu'il est entraîné sur de grandes quantités de données.

Par ailleurs, nous avons étudié les méthodes permettant de rendre un modèle multimodal. Nous avons consulté l'article [**Multimodal**], qui décrit plusieurs approches principales, notamment la fusion précoce (Early Fusion), la fusion tardive (Late Fusion) et la fusion intermédiaire (Intermediate Fusion) :

• Fusion précoce : Cette méthode consiste à combiner dès le début les données issues de différentes modalités, souvent après un prétraitement, avant de les transmettre au réseau de neurones.

• Fusion tardive : Chaque modalité est d'abord traitée séparément par un modèle spécialisé, puis les résultats sont combinés à la fin pour produire une prédiction finale.

• Fusion intermédiaire : Les données sont d'abord traitées séparément, puis fusionnées à un stade intermédiaire, avant la phase de prédiction.

Ces approches nous ont permis d'explorer des solutions adaptées à notre projet, en identifiant les avantages et les limites de chaque méthode en fonction de nos besoins.

Pour réaliser cette tâche, nous avons effectué des recherches supplémentaires, ce qui nous a conduits à explorer le concept du transfer learning. Cet article [**Transfer learning**] nous a introduit à cette notion. Cette méthode repose sur la réutilisation de connaissances déjà apprises par un modèle pré-entraîné, permettant ainsi d'optimiser les ressources nécessaires en évitant de repartir de zéro dans l'entraînement de nouveaux modèles. Elle offre également la possibilité d'ajuster ces modèles à des tâches spécifiques grâce au fine-tuning, garantissant des performances adaptées au contexte visé.

3 Description du travail réalisé

Notre projet s'est naturellement divisé en plusieurs tâches distinctes, et au sein de ces tâches, les membres du groupe n'ont pas tous été impliqués au même niveau. Cette partie a donc pour objectif de présenter le déroulement de notre projet, les différents points d'avancement ainsi que les personnes impliquées dans chaque tâche.

3.1 Mise en relation avec le client

Membres impliqués : tous.

La première étape d'un tel projet est de prendre connaissance des attentes du client. Pour ce faire, il est vital d'être en relation avec le client tout au long du projet. Nous avons donc rapidement contacté notre client, qui se trouve être professeur à CentraleSupélec, ce qui a facilité les rencontres. Nos premières discussions visaient à comprendre l'objectif global du projet, l'idée que le client se fait du produit à réaliser et le sujet sur lequel nous allions travailler. Notre client était très joignable et réactif par mail, ce qui a rendu les échanges très faciles, que ce soit en visioconférence ou en présentiel, pour lui poser des questions, lui demander des ressources ou simplement l'informer des avancées de notre projet. Cette tâche n'était donc pas limitée dans le temps et n'a conduit à aucun livrable, mais elle nous a accompagnés tout au long du projet et a été essentielle.

3.2 Lecture bibliographique

Membres impliqués : tous.

Une autre étape importante au démarrage du projet est le travail de renseignement bibliographique. Cette étape a été cruciale pour notre projet, d'abord parce que celui-ci se base en grande partie sur un projet déjà existant, et donc sur la connaissance de l'article scientifique associé. De plus, tout au long du projet, nous avons découvert de nouvelles méthodes d'intelligence artificielle, comme par exemple les transformers, qui étaient au cœur de notre projet. Nous avons également exploré des méthodes de traitement de signaux, d'implémentation de modèles multimodaux et d'autres points sur lesquels nous avons eu besoin de nous renseigner et de nous appuyer sur des articles scientifiques.

Cette tâche s'est beaucoup concentrée au début du projet mais a été nécessaire tout au long de son avancement. Elle n'est pas toujours évidente puisque tous les articles sont en anglais et contiennent des termes techniques et nouveaux pour beaucoup, notamment ceux qui proviennent de domaines très techniques du machine learning. De plus, les articles sont souvent très longs et très détaillés, ce qui ne facilitait pas la recherche d'informations.

3.3 Installation de l'environnement de travail

Membres impliqués : tous.

Notre client nous ayant informés des grandes quantités de calcul nécessaires à l'entraînement des modèles de fondation auxquels nous avions affaire, il nous a également fourni un accès au DCE de Metz. La première étape a donc été de se connecter à ce serveur distant à l'aide de nos identifiants. La connexion n'a pas posé de problème et, dans l'ensemble, le cluster a très bien fonctionné, à l'exception d'une panne survenue pendant nos créneaux de travail. De plus, l'espace disponible sur le cluster est partagé entre tous les utilisateurs (environ 7 To), et nous avons reçu à la fin de notre projet un message d'avertissement nous informant que nous prenions trop de place sur le cluster (environ 500 Go) et qu'il fallait libérer de l'espace. Cela n'a pas été très difficile : nous avons simplement supprimé les données et programmes dont nous n'avions plus besoin.

Ensuite, il a fallu installer le modèle SFM sur le serveur distant. Le projet possède un dépôt GitHub, ce qui nous a permis d'installer directement le programme via git clone. Cependant, plusieurs problèmes sont apparus lors de l'installation des différentes bibliothèques : il était nécessaire d'installer un environnement Conda et de manipuler manuellement l'installation de certaines librairies. Cette étape nous a bloqués relativement longtemps, car chaque fois qu'une erreur était réglée, une autre apparaissait. Mais nous avons réussi à surmonter les problèmes d'installation, et nous avons donc toutes les informations sur les différentes versions des bibliothèques à installer.

3.4 Entrainement de SFM

Membres impliqués : tous.

L'environnement de travail à distance installé, nous avons pu commencer à entraîner le modèle SFM sur différentes tâches. Cette partie a nécessité la manipulation de nouveaux outils, comme l'exécution de programmes via les commandes sbatch, qui permettent d'envoyer l'entraînement sur un GPU dédié pour de longues durées. Nous avons rapidement remarqué que l'entraînement d'un tel modèle prenait beaucoup de temps. Nous avons donc entraîné le modèle avec différentes tailles de batch et en faisant varier le nombre d'époques d'entraînement pour ensuite pouvoir évaluer l'influence de ces paramètres sur les performances du modèle. Mis à part le temps très important nécessaire à l'entraînement de ces modèles, cette étape n'a pas présenté de difficultés particulières et s'est étalée sur environ trois semaines.

3.5 Mise en place d'un outil de visualisation

Membres impliqués : Enio Yuhan.

Maintenant que nous avons entraîné le modèle sur les différentes tâches, il fallait vérifier la cohérence des résultats : voir si les résultats produits par le modèle semblent cohérents. Pour ce faire, nous avons implémenté un Jupyter Notebook pour visualiser la sortie du système sur une certaine tâche. Cet outil ne donne pour le moment aucune indication chiffrée sur les performances du modèle, mais il aide déjà à visualiser la cohérence des résultats renvoyés. L'implémentation de cet outil nous a pris une séance.

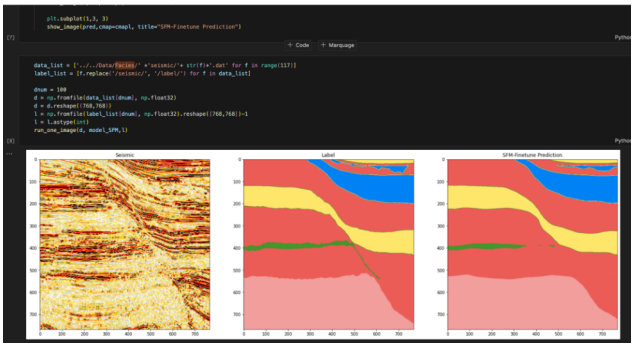

FIGURE 1 – Outil de visualisation

3.6 Mise en place d'une métrique pour estimer les performances du modèle

Membres impliqués : Yuhan.

Il nous fallait maintenant trouver un moyen de quantifier les performances de notre modèle pour les comparer à celles du papier. En effet, lors de la soutenance intermédiaire, on nous a fait remarquer qu'aucune donnée chiffrée n'était alors calculée pour comparer les performances de notre modèle à celles du papier. Pour cela, nous avons simplement repris les métriques utilisées dans le papier. Ces métriques sont l'IoU (Intersection over Union) et le CPA (Class Precision Accuracy).

L'IoU mesure la superposition entre la prédiction et la vérité terrain, ce qui est crucial pour évaluer la précision de la segmentation. Le CPA, quant à lui, évalue la précision de la classification des pixels, ce qui est essentiel pour vérifier que le modèle identifie correctement les différentes classes d'objets.

Les résultats obtenus sont très satisfaisants : les performances de notre modèle sont très proches, voire même légèrement meilleures que celles rapportées dans le papier. Cette amélioration peut s'expliquer par la configuration des GPU utilisés lors de l'entraînement, qui a permis une optimisation plus fine des paramètres du modèle.

Cette tâche n'a présenté aucune difficulté majeure et nous a pris environ deux semaines. Cette comparaison quantitative nous a permis de valider notre modèle.

3.7 SFM entraîner sur des signaux

Membres impliqués : Julien.

Nous nous sommes demander quelle serait une façon simple d'avoirun modèle multimodal. Nous avons donc eu l'idée de convertir les signaux que prendrait en général sisLM en des images de 224x224, dans l'objectif d'utiliser la tâche d'interpolation et ce à l'aide d'une transformation en spectrogramme de Mel.

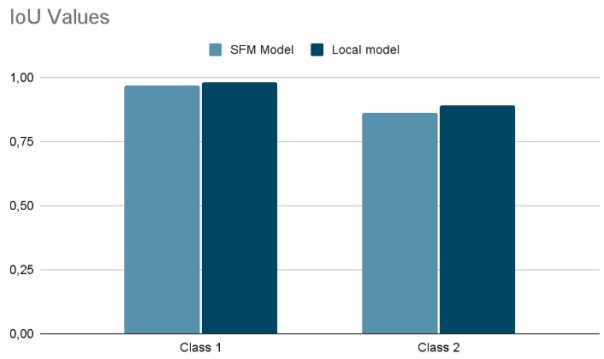

FIGURE 2 – Estimateur IoU de notre model VS celui du papier tache Salt

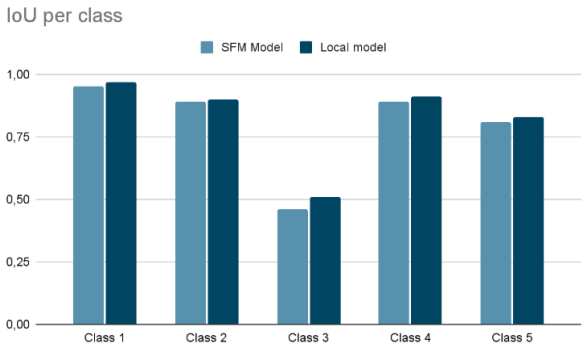

FIGURE 3 – Estimateur IoU de notre model VS celui du papier tache Facies

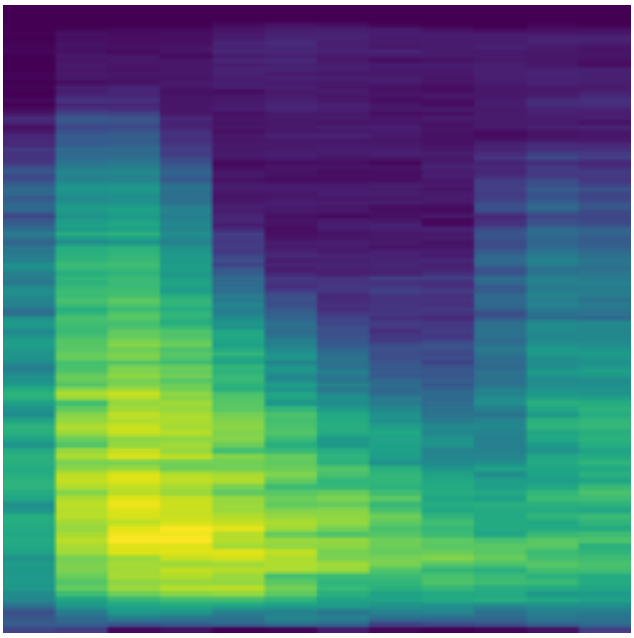

FIGURE 4 – Spectrogramme de Mel en 224x224

Pour permettre à SFM de traiter des signaux, nous avons commencé par transformer

ces derniers en images. Nous avons ensuite créé un jeu de données mixte composé à 50 % d'images issues des signaux (80 % de tremblements et 20 % de bruit) et à 50 % d'images du jeu de données original de SFM. Ce jeu de données mixte a été utilisé pour réentraîner le modèle, avec pour objectif de créer un modèle capable de prendre en entrée à la fois des signaux et des images.

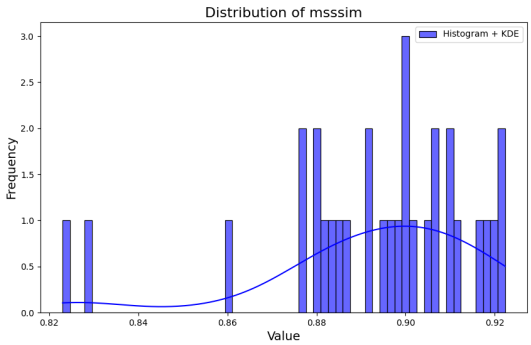

FIGURE 5 – Estimateur Msssim du nouveau modèle

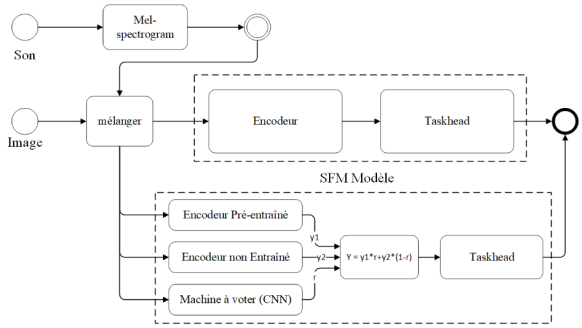

FIGURE 6 – Architecture

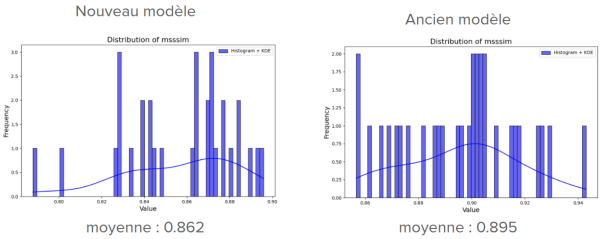

FIGURE 7 – Comparaison des deux modèles

3.8 Recherche d'un nouveau modèle pour la multimodalité

Membres impliqués : Timothée Julien Yuhan.

Après toutes ces tâches, il nous fallait maintenant réaliser un modèle multimodal qui pourrait être entraîné sur des données qui étaient à la fois visuelles (images prises par SFM) et audio (signaux 3D pris par seisLM). Nous nous sommes donc penchés sur la question de comment créer un modèle multimodal, et nous sommes tombés sur ce papier : [**Multimodal**]. Nous avons décidé de réaliser une early fusion, la percevant comme étant la solution la plus rapide à mettre en place. Nous nous sommes ensuite renseignés sur le transfer learning grâce à des articles, mais notamment grâce à un TP fait par M. Hervé Le Borgne qui nous a permis de mieux nous approprier la notion qu'avec la simple lecture d'articles.

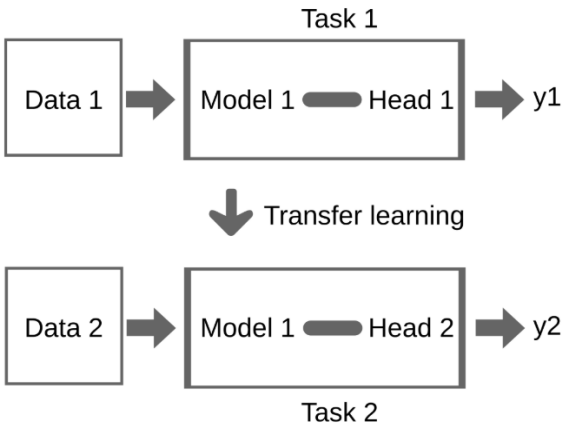

FIGURE 8 – Transfer learning

Après avoir complété ce TP, nous nous sommes lancés dans la conception de notre modèle. Pour ce faire, nous voulions transférer les données d'entraînement de SFM et de seisLM sur un modèle déjà multimodal. Notre dévolu s'est jeté sur ImageBind, un modèle fait par Meta qui semblait adéquat à nos demandes. La stratégie était la suivante : remplacer les encodeurs audio et visuels d'ImageBind par ceux de seisLM et de SFM respectivement. De cette façon, nous n'avions pas à nous soucier de problèmes de typage de données. Nous ajoutions ensuite quelques couches de fusion, pour commencer de la même taille, mais nous nous sommes dit qu'il serait peut-être plus judicieux de les faire en taille décroissante. À la fin du modèle, nous mettions une couche de classifieur pour transformer notre vecteur de probabilité en une prédiction.

Nous avons également dû faire d'autres choix, comme par exemple celui de l'optimiseur. Au début, nous sommes partis sur la fonction "CrossEntropy", comme utilisée dans le TP de M. Le Borgne, mais nous avons finalement fait le choix d'utiliser la fonction "NLLLoss" qui a déjà un soft-max implémenté.

Nous avons décidé pour l'instant de faire un travail de classification. Le modèle pourra faire deux tâches séparées : classifier parmi les classes de SFM, ou bien celles de seisLM.

Nous avons ensuite dû chercher comment implémenter les encodeurs, trouver les datasets d'entraînement, ainsi que gérer ces données. Cela a été bien plus compliqué que prévu, n'étant que partiellement documenté dans les projets.

Nous avons rencontré de nombreuses difficultés lors de cette étape. Certaines ont été surmontées, d'autres non, dues à un manque de temps. Nous avions, entre autres, des problèmes avec l'importation des modules d'ImageBind dans le programme.

La difficulté majeure qui a été rencontrée était le manque de datasets pour entraîner le modèle à la fois sur des images et des signaux, ce qui avait été imaginé lors de la conception du code, mais a dû être changé vers la fin dans la tentative de l'entraîner comme décrit plus haut.

4 Conclusion et Perspectives

Tout au long de ce projet, nous avons exploré divers modèles de fondation, tels que SFM et SeisLM, et avons réussi à nous les approprier, ce qui nous a permis de progresser sur deux aspects essentiels. D'une part, nous avons approfondi notre compréhension des modèles de fondation, en nous familiarisant avec leurs spécificités et en apprenant à les adapter à notre problématique. D'autre part, nous avons renforcé nos compétences techniques, en développant notre maîtrise des outils nécessaires à l'utilisation de notre modèle. Nous avons acquis de nombreuses connaissances et produit deux résultats concrets : un code utile pour la suite du projet, ainsi qu'un modèle SFM entraîné sur le son, capable de traiter à la fois des signaux et des images, bien que pas de manière simultanée. Ces avancées nous permettent de mieux appréhender les défis à venir et de solidifier notre approche technique. Ce projet a également été l'occasion pour nous d'approfondir nos connaissances et compétences sur des sujets de recherche, à travers une veille documentaire sur les articles scientifiques en lien avec notre sujet, ainsi qu'une réflexion sur la manière d'amener un projet fortement lié à des domaines de recherche. D'un point de vue moins technique, les spécificités de notre groupe ont été une occasion de faire preuve de qualités importantes dans la réalisation de projets, telles que le travail d'équipe, illustré par l'inclusion de membres internationaux, ainsi que l'organisation et la communication.

Concernant les perspective de notre projet nous pensons qu'il faudrait notamment approfondir la piste du modèle multimodal se basant sur le modèle ImageBind de Meta. Cette piste est intéressante, car elle permettrait de corréler les différentes modalités, et l'on pourrait envisager d'ajouter également des données textuelles, ce qui pourrait être très intéressant.

Références

[1] Hanlin Sheng, Xinming Wu, Xu Si, Jintao Li, Sibo Zhang, and Xudong Duan, Seismic Foundation Model (SFM) : a new generation deep learning model in geophysics, 15 Dec 2023.

[2] Tianlin Liu, Laura Laurenti, Jannes Münchmeyer, Chris Marone, Ivan Dokmani, SeisLM : a Foundation Model for Seismic Waveforms, 21 Oct 2024.

[3] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, Multimodal Foundation Models : From Specialists to General-Purpose Assistants, 18 Sep 2023.

[4] Yuxuan Liu, Jingmin Sun, Xinjie He, Griffin Pinney, Zecheng Zhang, Hayden Schaeffer, PROSE-FD : A Multimodal PDE Foundation Model for Learning Multiple Operators for Forecasting Fluid Dynamics 15 septembre 2024

[5] Mehra, S., Ranga, V. and Agarwal, R. (2024), Multimodal Integration of Mel Spectrograms and Text Transcripts for Enhanced Automatic Speech Recognition : Leveraging Extractive Transformer-Based Approaches and Late Fusion Strategies. Computational Intelligence, 40 : e70012. https ://doi.org/10.1111/coin.70012

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby : Transformers for image recognition at scale.

[7] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Senior Member, IEEE, Hui Xiong, Fellow, IEEE, and Qing He

[8] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, Shu Zhang : A Comprehensive Review of Multimodal Large Language Models : Performance and Challenges Across Different Tasks

[9] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Senior Member, IEEE,Hui Xiong, Fellow, IEEE, and Qing He : A Comprehensive Survey on Transfer Learning