

Stage M2 recherche : Optimisation du temps d'inférence des *Foundation Models* pour la prévision de séries temporelles

Ubiquity AI - Bordeaux - 6 mois

Contexte

La prévision de séries temporelles (*Time Series Forecasting*) ou (*TSF*) est un élément clé de l'analyse de données moderne. Elle trouve des applications dans de nombreux domaines tels que la finance, la santé, l'économie ou encore la prévision météorologique [1].

À l'instar des modèles GPT en traitement du langage naturel, les *Foundation Models (FM)* ont révolutionné la pratique de *TSF*.

De la même manière que GPT apprend les structures linguistiques générales à partir de textes à grande échelle, ces modèles apprennent les dépendances temporelles à partir de données séquentielles diverses.

Leur pré-entraînement leur permet de capturer des schémas universels tels que les tendances, les cycles et la saisonnalité.

Cependant, leur grande taille et leurs architectures complexes peuvent engendrer une latence d'inférence élevée [2], notamment lorsqu'ils sont déployés sur des CPUs, ce qui limite leur utilisation dans des applications en temps réel ou à ressources limitées.

Ce stage a pour objectif de considérer le problème du temps d'inférence de ces modèles lorsqu'ils sont déployés sur des CPUs, et essayer d'optimiser ce temps afin de rendre ces modèles plus accessible à un grand spectre d'applications réelles.

À propos d'Ubiquity AI

Ubiquity AI est une deeptech bordelaise, essaimage du LIPN-CNRS de l'Université Sorbonne Paris Nord avec plus de 20 ans de recherche fondamentale et 10 ans de R&D travaillant sur le stream learning. Sa mission est de démocratiser l'intelligence artificielle en la rendant accessible grâce au stream learning qui est la forme la plus pure et naturelle d'apprentissage.

L'objectif

Ce stage s'inscrit dans le cadre d'un projet plus large portant sur le stream learning dans les modèles de *TSF* basés sur les *FM*s.

Son objectif principal est d'étudier le compromis temps d'inférence / performance prédictive et d'explorer les facteurs qui peuvent y jouer un rôle principal dans un premier temps. En deuxième temps, explorer la possibilité d'optimisation du temps d'inférence en exploitant les derniers jeux d'instructions CPU. La première étape de ce travail consiste à créer un benchmark similaire à celui présenté dans [2], mais appliqué aux CPUs, afin d'évaluer l'état de l'art dans ce domaine.

Les tâches principales

- Réaliser une revue bibliographique sur le *TSF* en utilisant les *FM*s.
- Étudier les *FM*s les plus populaires pour le *TSF* et apprendre à les utiliser correctement.
- Créer un répertoire de jeux de données divers pour le *TSF*.
- Produire un *benchmark* de ces modèles sur les jeux de données sélectionnés en utilisant les métriques pertinentes (temps, CPU, mémoire).
- Étudier les approches possibles pour optimiser le temps d'exécution sur CPU de ces modèles, en exploitant les jeux d'instructions CPU les plus récentes.
- Réaliser une étude comparative des modèles avant et après optimisation.

Ce que vous allez apprendre

- Les bases du *TSF* avec des *FM*s.
- La conduite d'un travail bibliographique sur un sujet spécifique : *TSF* à l'aide des *FM*s.
- La collecte de jeux de données pertinents.
- L'application de tests statistiques pour comparer les performances des modèles.
- L'optimisation du temps d'exécution sur CPU d'un algorithme donné.
- L'utilisation des *FM*s les plus populaires pour le *TSF* (TimesFM [4], Chronos [5], ... etc.).
- L'utilisation et l'extension des bibliothèques Python les plus connues pour le *TSF*, telles que Sktime.

Plus de détails

- **Lieu** : Bordeaux centre-ville.
- **Période** : Février-Mars à Août-Septembre 2026 (6 mois).
- **Niveau** : Master 2 (ou équivalent ingénieur).
- **Gratification** : 1000€ brut par mois.

- **Profil recherché** : le candidat doit avoir de l'intérêt pour l'optimisation software/ hardware, l'architecture des CPU modernes, l'efficacité énergétique de l'IA.
- **Candidature** : Envoyez votre CV à l'adresse suivante : subhy.albakour@idaaas.com
- **La suite** : Le stage peut être suivi par une thèse CIFRE pour continuer le travail sur l'apprentissage continu de modèles de prévision de séries temporelles qui sera plus adapté aux applications réelles vu la nature dynamique des données ces applications [3]. Pour ce poste, il est attendu un bon niveau en mathématique, en algorithmique et en programmation.

Références

- [1] S. R. K. Kottapalli et al., “Foundation Models for Time Series: A Survey,” Apr. 05, 2025, arXiv: arXiv:2504.04011. doi: 10.48550/arXiv.2504.04011.
- [2] F. Montet, B. Pasquier, and B. Wolf, “Benchmarking Foundation Models for Time-Series Forecasting: Zero-Shot, Few-Shot, and Full-Shot Evaluations,” Computer Sciences & Mathematics Forum, vol. 11, no. 1, p. 32, 2025, doi: 10.3390/cmsf2025011032.
- [3] P. Gong, E. Eldele, M. Wu, Z. Chen, X. Li, and D. Zhang, “Bridging Distribution Gaps in Time Series Foundation Model Pretraining with Prototype-Guided Normalization,” Apr. 15, 2025, arXiv: arXiv:2504.10900. doi: 10.48550/arXiv.2504.10900.
- [4] A. Das, W. Kong, R. Sen, and Y. Zhou, “A decoder-only foundation model for time-series forecasting,” Apr. 17, 2024, arXiv: arXiv:2310.10688. doi: 10.48550/arXiv.2310.10688.
- [5] A. F. Ansari et al., “Chronos: Learning the Language of Time Series,” Nov. 04, 2024, arXiv: arXiv:2403.07815. doi: 10.48550/arXiv.2403.07815.