

ANTONIO SOGGIA

antoniosoggia5@gmail.com • antoniosoggia.it • github.com/antoniosoggia • FR : +33 7 49 91 79 66

Profil

Docteur en Machine Learning et Traitement du Signal avec une solide expertise en génie logiciel. Spécialisé dans le développement de CNN légers pour les environnements bruités et la construction d'infrastructures IA auto-hébergées. Entièrement autonome sur l'ensemble du cycle ML : de la modélisation mathématique et la génération de données synthétiques au déploiement backend et à l'optimisation multi-GPU.

Compétences Techniques

- **Langages de Programmation** : Python, C/C++, SQL, Bash, Scala
- **Machine Learning et Mathématiques** : PyTorch, TensorFlow, CNN, LLM (Ollama, llama.cpp), Traitement du Signal (DSP), Régression Symbolique, Algèbre Linéaire, Optimisation
- **MLOps et Ingénierie** : Docker, Linux/SSH, Flask, API REST, Spark, Git, Optimisation multi-GPU, JAX, MoE

Expérience Professionnelle

AI Scientist (anciennement R&D Apprentice) - Thales DMS

Sophia Antipolis (FR) | 2021 - 2026

- **Développement ML de bout en bout** : Piloté la transition d'un pipeline de traitement du signal sonar depuis le prototype R&D jusqu'à un système breveté prêt pour la production.
- **Simulation et Data Engineering** : Conçu un simulateur sonar basé sur la physique pour générer des jeux de données synthétiques à grande échelle, résolvant un manque critique de données réelles pour l'entraînement ML.
- **Conception Architecturale** : Développé *AntoNet*, un CNN léger personnalisé pour le débruitage de signaux 3D. Atteint un gain de 11 dB (proche de la limite théorique) avec seulement 1.2k paramètres, optimisé pour l'inférence temps réel sur GPU grand public.
- **Analyse Mathématique** : Conduit une analyse du flux de gradient par régression symbolique pour caractériser la variété des minimiseurs globaux, démontrant la robustesse des architectures linéaires face au bruit gaussien.

Projets d'Ingénierie Sélectionnés

Serveur IA Privé et Infrastructure LLM

Docker, Ollama, CUDA, Tailscale

- Conçu un serveur multi-GPU (RTX 4070+3060) déployant des modèles MoE quantifiés via Ollama et llama.cpp ; implémenté l'allocation automatique de la VRAM et l'accès distant via VPN Zero-Trust Tailscale.
- Développé un moteur de recherche web auto-hébergé et dockerisé pour la Web-Augmented Generation (WAG) sans API externe.

Orchestra - Plateforme de Traduction IA Locale

Python, Flask, Ollama, System Design

- Conçu un pipeline NLP hybride utilisant une architecture déterministe lookup-first ; le système résout la terminologie confirmée depuis une mémoire sémantique avant l'inférence LLM, garantissant 100% de cohérence terminologique.
- Implémenté l'Analyse de Topologie Documentaire pour la déduplication des paragraphes, obtenant une réduction de 85% des appels LLM ; renforcé la fiabilité via la détection de dérive linguistique et les gardes d'alignement de paragraphes.
- Construit un Store Sémantique Persistant fonctionnant comme une Mémoire de Traduction locale, réduisant la dépendance à l'inférence à mesure que la couverture du domaine de l'utilisateur croît au fil du temps.

Publications et Brevets

- **Brevet Déposé** : Pipeline de Débruitage et Suivi 3D;
- **Article de Revue** : *AntoNet: A Small Linear Contrast-Enhancing CNN for Pixel-Wise Detection*. (En révision, Signal Processing: Image Communication). Invité après un classement Top-5 à IPTA 2024;
- **Articles de Conférence** : *AntoNet* (IPTA 2024) ; *Lightweight CNN for Sonar Detection* (EUSIPCO 2025).

Formation

Doctorat en Mathématiques Appliquées et Machine Learning, Université Côte d'Azur & Thales DMS

05/2026

Master en Ingénierie Mathématique, Université Côte d'Azur

2022, mention très bien

Licence en Mathématiques Appliquées (MASS), Université Côte d'Azur

2020, mention très bien

Enseignement et Langues

Enseignement : Valorisation des Données (Apprentissage supervisé, ROC/AUC) et Big Data (SQL/Spark) - Chargé de TP à Polytech UCA.

Langues : Français (Langue maternelle), Italien (Langue maternelle), Anglais (Courant).