NADIR TRAPSIDA

trapsidanadir@gmail.com | 819 580-9669 | Montréal, QC, Canada
GitHub | LinkedIn | Portfolio | Medium | LinkTree

Résumé

Ingénieur en IA/ML avec une formation en génie informatique et électronique, spécialisé dans la mise en production à grande échelle de modèles d'apprentissage automatique. Expérience avérée dans le déploiement de pipelines ML de bout en bout, l'optimisation de systèmes d'inférence traitant plus de 150K requêtes quotidiennes, et la création de produits utilisés par les grandes flottes nord-américaines.

Expérience professionnelle

Ingénieur en apprentissage automatique *Sep 2024 – Présent*
Isaac Instruments Montréal, QC

- Réduction de la latence d'inférence de l'IA de 50% (80 ms → 40 ms) pour 150K+ requêtes quotidiennes en réorganisant l'infrastructure de déploiement des modèles et en optimisant la couche API, réduisant le temps moyen de traitement des requêtes de bout en bout de 33%.
- Conception et déploiement d'un pipeline d'analyse vidéo de bout en bout traitant plus de 7K vidéos/heure via 20 modèles ML, en collaboration avec l'équipe DevOps pour la surveillance intégrée, la planification efficace des ressources et la récupération automatique, permettant de passer du prototype à la production.
- Développement d'un assistant vocal TTS/STT utilisant OpenWakeWord pour les conducteurs de camions sur du matériel edge à ressources limitées, permettant une interaction mains libres pour améliorer l'expérience et la sécurité des conducteurs.
- Identification d'une lacune critique dans la solution de numérisation existante, puis collaboration avec l'équipe produit pour développer un POC et le transformer en une fonctionnalité de production améliorant la numérisation et l'extraction de données sur tablette pour tous les clients d'Isaac.
- Création d'une expérience "Driver Year in Review" avec des éléments de gamification (statistiques, badges, jalons) pour améliorer la rétention et l'engagement des conducteurs dans les flottes partenaires.
- Accélération de la création de jeux de données pour plusieurs équipes en développant un outil interne d'annotation d'images/vidéos, augmentant le débit d'étiquetage et réduisant les cycles de révision.
- Livraison d'une application d'événements de sécurité utilisée par les grandes flottes canadiennes et américaines en développant le backend FastAPI et en contribuant au frontend Next.js, offrant aux clients un accès en temps réel aux informations de sécurité.
- Prototypage d'un chatbot interne de recherche de connaissances basé sur RAG avec récupération hybride BM25 + vecteur, améliorant la découverte de documents pour les équipes d'ingénierie et d'opérations.

Cofondateur et Directeur de l'IA *2022 – 2024*
Queva Inc. Montréal, QC

- Architecture et déploiement de l'un des premiers agents conversationnels d'assistance vétérinaire alimentés par RAG, augmentant l'engagement des utilisateurs de 35% et améliorant la rétention des clients.
- Conception du pipeline de déploiement en production pour le système RAG, atteignant une disponibilité de 99,5% et réduisant les erreurs d'inférence dans une application grand public sensible à la latence.
- Entraînement d'un modèle prédictif sur des données de capteurs IMU pour classifier les activités des chiens avec une précision de 97,5%, permettant des analyses comportementales en temps réel pour les propriétaires d'animaux.
- Développement du backend IoT intégrant base de données, application mobile et services d'inférence ML, réduisant de moitié le délai d'intégration initial.
- Gestion et optimisation de l'infrastructure multi-services sur Microsoft Azure, réduisant les coûts mensuels de cloud de 30% tout en améliorant l'évolutivité.
- Optimisation des pipelines CI/CD, réduisant le temps de déploiement de 3 heures et augmentant la fiabilité des versions pour l'équipe d'ingénierie.
- Conception du prototype d'application mobile ayant conduit à un pitch réussi, obtenant un investissement de 250K$ auprès de VCs.
- Gestion d'une équipe interfonctionnelle de 6 développeurs et stagiaires, améliorant le rythme de livraison des projets et la productivité de l'équipe.

Ingénieur de recherche en IA *2022*
Laboratoire d'Intelligence Véhiculaire, Université de Sherbrooke (LIV) Sherbrooke, QC

- Développement d'une simulation 3D multi-caméras d'un bus autonome dans CARLA, permettant des tests reproductibles des algorithmes de perception et de planification.
- Amélioration de 13% de la précision de détection d'un modèle de détection d'objets sur le simulateur CARLA, renforçant la validation de sécurité pour la recherche sur la conduite autonome.
- Restructuration et intégration de plusieurs modules d'une grande base de code de recherche sur les véhicules autonomes, accélérant les itérations expérimentales et améliorant la maintenabilité.
- Collaboration avec NovaBus pour préparer l'intégration des algorithmes validés en simulation sur une plateforme de bus physique.

Développeur logiciel en IA *2021*
Centre de Recherche Informatique de Montréal (CRIM) Montréal, QC

- Optimisation des workflows CI/CD avec Makefiles et GitHub Actions, réduisant les cycles de déploiement et de test.
- Augmentation de 20% de la couverture de code grâce au développement systématique de tests unitaires, améliorant la fiabilité de la base de code.
- Découverte d'une faille de sécurité critique dans un produit en production, prévenant des potentielles violations de données et protégeant la vie privée des utilisateurs.

Data Scientist et Ingénieur en vision par ordinateur *2020 – 2021*
Decathlon Inc. Montréal, QC

- Développement d'un modèle de classification de texte avec une précision de 95% pour distinguer le contenu lié à Decathlon des mentions sportives génériques, permettant une surveillance évolutive de la marque sur les plateformes sociales.
- Optimisation des points de terminaison de service ML avec Protocol Buffers, réduisant la latence des prédictions de 83% (12 s $\rightarrow$ 2 s) et améliorant le débit pour les consommateurs en aval.
- Création d'une application de recherche d'images de bout en bout (frontend, backend, base de données) utilisant des embeddings de vecteurs de mots, atteignant 90% de pertinence pour la récupération d'images sur les réseaux sociaux.

Spécialiste en intégration de solutions matérielles et logicielles, R&D *2018 – 2019*
Kontron Inc. Boisbriand, QC

- Automatisation des benchmarks de performance réseau et de consommation d'énergie à l'aide de Python et Bash, réduisant le temps de test de 4 heures par cycle et accélérant la qualification du matériel.
- Réalisation de benchmarks GPU (NVIDIA Tesla P4, T4 ; AMD Vega 4000) sur 5 familles de serveurs, garantissant la compatibilité des pilotes et les performances de base pour les déploiements clients.

Compétences techniques

ML & IA : PyTorch, TensorFlow, Keras, Scikit-Learn, ONNX, OpenCV, MLflow, Hugging Face, LangChain, LlamaIndex, LLMs (GPT, LLaMA), Systèmes RAG, Ingénierie de prompts, LoRA Fine-tuning, NumPy

Backend & Données : Python, FastAPI, Flask, gRPC, PostgreSQL, MongoDB, MSSQL, Redis, Pandas, PyTest

Frontend : JavaScript, Vue.js, Next.js

Cloud & Infrastructure : AWS (S3, SageMaker, Lambda, ...), Azure (Container Apps (Kubernetes-managed), IoT Hub, DevOps, Functions, Blob Storage, Web Apps, VMs, ...), OVH, Docker, Grafana

DevOps : Git, GitHub Actions, Pipelines CI/CD, Bash

Langues : Français (natif), Anglais (courant)

Formation

Université de Sherbrooke Sherbrooke, QC
Maîtrise en génie informatique, spécialisation en intelligence artificielle Bourse d'excellence académique
Baccalauréat en génie informatique, programme coopératif

Projets et publications

- **DocTrack.io** (Cofondateur) : Développement d'une plateforme SaaS B2B automatisant la classification, le renommage et le suivi des dates d'expiration des documents pour les entreprises de gestion immobilière, en utilisant des pipelines d'ingestion et d'extraction basés sur des LLM.

- **Déploiement de PyTorch avec FastAPI pour gérer 120K+ requêtes par jour** : Publié dans Data Science Collective, détaillant l'architecture de déploiement ML en production et les stratégies d'optimisation.
- **Nous sommes passés de TensorFlow à PyTorch et voici ce qui s'est passé** : Publié dans Level Up Coding, partageant des leçons tirées d'une migration réelle d'un framework ML en production.
- **Traduction LLM pour langues à faibles ressources** : Affinage d'un LLM avec LoRA pour traduire l'anglais en Zarma/Songhai, une langue africaine non prise en charge par Google Translate. Publié dans Data Science Collective et Towards AI.