Optimisation et Compression de Modèles IA pour le Déploiement Edge : Quantization, Pruning, Distillation

Lien : https://innov-systems.com/formation/optimisation-et-compression-de-modeles-ia-pour-le-deploiement-edge-quantization-pruning-distillation

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI145

CATÉGORIE
Edge AI, TinyML et IA Embarquée

Innov Systems

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Comprendre les contraintes du déploiement Edge AI
- Maîtriser les techniques de quantization, pruning et distillation
- Réduire la taille et la latence des modèles IA
- Optimiser les performances sur dispositifs embarqués
- Adapter les architectures de réseaux neuronaux
- Déployer des modèles IA en environnement contraint
- Améliorer l’efficacité énergétique des systèmes IA
- Utiliser les frameworks d’optimisation edge
- Mettre en place des pipelines de déploiement robustes
- Assurer la maintenance des modèles en production
- Étudier des cas concrets de Edge AI
- Développer des solutions IA adaptées à l’embarqué

POUR QUI ?

- Machine learning engineers
- Ingénieurs IA embarquée
- Data scientists
- Développeurs logiciels
- Architectes systèmes
- Ingénieurs IoT
- DevOps IA
- Consultants en intelligence artificielle
- Chefs de projet IA
- Product managers techniques
- Toute personne impliquée dans le déploiement de modèles IA sur edge devices

Innov Systems

Programme détaillé

1 / Introduction à l'IA embarquée (Edge AI)

- Définition du Edge AI et contraintes associées
- Différence entre cloud et edge computing
- Enjeux de latence, mémoire et consommation

2 / Contraintes de déploiement des modèles IA

- Limitations matérielles des dispositifs edge
- Besoins en temps réel et faible latence
- Arbitrages performance vs ressources

3 / Quantization des modèles IA

- Réduction de la précision des poids
- Types de quantization (post-training, quantization-aware training)
- Impact sur la performance et la taille du modèle

4 / Pruning des réseaux neuronaux

- Suppression des connexions non essentielles
- Pruning structuré vs non structuré
- Optimisation de la sparsité des modèles

5 / Knowledge distillation

- Transfert de connaissance entre modèles
- Teacher–student models
- Réduction de complexité tout en conservant la performance

Innov Systems

6 / Optimisation des architectures de modèles

- Sélection de modèles légers (MobileNet, EfficientNet)
- Adaptation des architectures pour l’edge
- Trade-off précision / efficacité

7 / Accélération matérielle et frameworks edge

- GPU, TPU et NPU embarqués
- Frameworks d’optimisation (TensorRT, ONNX Runtime)
- Compatibilité des modèles optimisés

8 / Pipeline de déploiement Edge AI

- Préparation des modèles pour production
- Conversion et export des modèles
- Intégration dans les dispositifs embarqués

9 / Monitoring et maintenance des modèles edge

- Suivi des performances en environnement réel
- Mise à jour des modèles déployés
- Gestion des dérives de données

10 / Études de cas et implémentations pratiques

- Déploiement IA sur mobile et IoT
- Optimisation de modèles réels
- Bonnes pratiques industrielles

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

16 au 20 Nov. 2026	Distanciel
12 au 18 Jan. 2027	Présentiel - Casablanca
08 au 12 Mars 2027	Présentiel - Casablanca
03 au 07 Mai 2027	Présentiel - Casablanca
28 Juin au 02 Juil. 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés