

Mis à jour le 14/10/2026

S'inscrire

Formation TensorRT avec Ultralytics YOLO

2 jours (14 heures)

Présentation

Associez TensorRT et Ultralytics YOLO pour préparer des applications de détection adaptées aux contraintes de production. Explorez les précisions FP32, FP16 et INT8 ainsi que les paramètres de compilation pour rechercher le meilleur équilibre entre vitesse, qualité des prédictions et consommation mémoire sur GPU.

Notre formation TensorRT avec Ultralytics YOLO vous permettra de passer d'un modèle entraîné à un moteur d'inférence optimisé, en comprenant les choix techniques qui conditionnent son utilisation sur votre matériel cible.

Vous apprendrez à préparer une référence de mesure, à maîtriser la chaîne ONNX et TensorRT et à configurer les dimensions d'entrée, les lots et la mémoire de construction. Vous expérimenterez également la quantification INT8 avec des données représentatives pour évaluer ses effets sur la détection.

À l'issue de la formation, vous serez en mesure de comparer les configurations selon la latence, le débit et la mAP puis d'intégrer le moteur retenu dans un pipeline de traitement d'images ou de vidéos.

Cette formation aborde également le profilage et le diagnostic des erreurs, ainsi que la reproductibilité des exports et les contraintes de compatibilité, afin de préparer des déploiements fiables et mesurables.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Préparer une référence de qualité et de performance pour un modèle de détection YOLO
- Exporter un modèle Ultralytics YOLO vers un moteur TensorRT adapté au matériel cible
- Configurer les dimensions d'entrée, les lots et les précisions de calcul
- Évaluer les compromis entre latence, débit, mémoire et qualité des détections
- Intégrer et valider un moteur optimisé dans un pipeline d'inférence reproductible

Public visé

- Ingénieur IA
- Ingénieur en vision par ordinateur
- Data Scientist
- Développeur Python
- Ingénieur MLOps

Pré-requis

- Pratique de Python pour écrire et exécuter des scripts de traitement de données
- Connaissance des tenseurs et de l'inférence avec PyTorch ou expérience équivalente
- Première utilisation de Ultralytics YOLO pour la prédiction ou la validation d'un modèle
- Compréhension des principes de détection d'objets et des métriques de précision et de rappel

Pré-requis techniques

- Poste sous Ubuntu 22.04 LTS ou Ubuntu 24.04 LTS avec droits d'installation, ou accès distant à un environnement équivalent
- Configuration recommandée pour les ateliers : processeur 4 cœurs, 16 Go de RAM, 30 Go disponibles et GPU NVIDIA Turing ou plus récent avec 8 Go de mémoire vidéo
- Environnement Python 3.10 avec PyTorch et Ultralytics dans des versions figées et validées avant la session
- Pilote NVIDIA et pile CUDA et TensorRT compatibles, avec trtexec installé via un paquet système, une archive ou un conteneur validé
- Éditeur Visual Studio Code ou équivalent et connexion Internet stable autorisant le téléchargement des dépendances, modèles et jeux de données

Programme de notre formation TensorRT avec Ultralytics YOLO

[Jour 1 - Matin]

Préparer le modèle et les mesures de référence

- Positionnement de TensorRT dans le cycle de déploiement d'un modèle Ultralytics YOLO
- Vérification de l'environnement : GPU, pilote NVIDIA, CUDA et dépendances
- Anatomie du pipeline de détection : prétraitement, inférence et post-traitement

- Préparation des images et annotations, distinction entre données de calibration et jeu d'évaluation indépendant
- Construction d'une référence PyTorch : qualité des détections, latence, débit et mémoire GPU
- Atelier pratique : Charger un modèle YOLO préentraîné, exécuter une validation et consigner les mesures de référence sur le GPU cible.

[Jour 1 - Après-midi]

Exporter et construire un moteur TensorRT

- Chaîne de conversion PyTorch, ONNX et moteur TensorRT : graphe, opérateurs et artefacts
- Export avec Ultralytics : format engine, résolution, taille de lot et budget mémoire de construction
- Dimensions statiques et dynamiques : profils de dimensions minimales, optimales et maximales
- Choix des précisions FP32 et FP16 selon le matériel et les résultats de validation
- Gestion des sorties : coordonnées, scores, classes et NMS selon l'architecture du modèle
- Atelier pratique : Construire des moteurs FP32 et FP16, les charger avec Ultralytics et comparer leurs prédictions à la référence.

[Jour 2 - Matin]

Quantifier et mesurer les compromis

- Principes de quantification INT8 et effets possibles sur les scores et la localisation des objets
- Préparation d'une calibration représentative : diversité des scènes, objets et conditions d'acquisition
- Paramétrage de l'export quantifié et lecture des journaux selon la version de TensorRT
- Protocole de benchmark avec trtexec : préchauffage, répétitions, synchronisation et transferts CPU-GPU
- Comparaison FP32, FP16 et INT8 : mAP, latence, débit et consommation mémoire à conditions identiques
- Atelier pratique : Exporter un moteur INT8 avec calibration, mesurer sa qualité sur le jeu indépendant et sélectionner une configuration selon des critères d'acceptation.

[Jour 2 - Après-midi]

Intégrer et fiabiliser le déploiement

- Intégration du moteur dans un pipeline Python de détection sur images et vidéos
- Mesure de la latence de bout en bout et identification des coûts de décodage, prétraitement et post-traitement
- Diagnostic des erreurs de dimensions, mémoire, opérateurs et désérialisation
- Contraintes de compatibilité des moteurs entre GPU, plateformes et versions, stratégie de reconstruction
- Reproductibilité du déploiement : versions figées, paramètres d'export, tests de non-régression et suivi des performances

- Atelier pratique : Livrer un pipeline vidéo utilisant le moteur retenu, vérifier les critères de qualité et de latence puis documenter sa reconstruction.

Sociétés concernées

Cette formation s'adresse à la fois aux particuliers ainsi qu'aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d'auto-évaluation nous permettant d'apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d'anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l'acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.