Résumé

Conception et déploiement distribué de modèles IA pour des applications de vision par ordinateur et d'industrie 4.0

Auto DIST-Framework: *Automatic, DIstributed and Synchronous Training*

Cette thèse s'inscrit dans le contexte de l'industrie 4.0, où l'intégration de l'intelligence artificielle, notamment du deep learning, transforme les processus industriels. Cependant, l'entraînement des modèles IA dans des environnements à ressources limitées reste un défi majeur, particulièrement pour les petites et moyennes entreprises. Pour répondre à ce besoin, un framework nommé Auto-DIST (Automatic Distributed and Synchronous Training) a été conçu.

Auto-DIST propose une approche innovante pour distribuer et optimiser l'entraînement des modèles IA sur des infrastructures modestes, en exploitant efficacement les ressources disponibles. Il s'appuie sur des mécanismes avancés de surveillance, de prédiction, d'optimisation et de distribution, permettant de surmonter les contraintes liées à la complexité croissante des modèles. Ce framework a été testé sur des cas d'utilisation concrets, tels que la maintenance prédictive et le contrôle qualité, dans le cadre de projets industriels, où il a démontré sa capacité à réentraîner rapidement des modèles tout en maintenant une haute précision.

Les expérimentations ont d'abord été menées sur une infrastructure modeste composée de plusieurs types de machines. Sur l'un de ces types de machines, un total de 23 544 expériences a été réalisé afin d'évaluer finement le comportement du framework dans des conditions réalistes. Pour valider la généricité et la robustesse des résultats, les expériences ont ensuite été reproduites sur l'infrastructure haute performance Grid 5000, en utilisant l'ensemble des machines auxquelles l'accès a été possible.

Les résultats montrent qu'Auto-DIST peut réduire significativement le temps d'entraînement en exploitant au mieux les ressources disponibles. Le framework a su s'adapter au matériel et au nombre de machines disponibles sur les différentes infrastructures, démontrant ainsi sa scalabilité et son élasticité. Selon les cas, il a permis d'atteindre un facteur d'accélération allant de 1,124× sur CPU et 1,407× sur GPU dans les scénarios les plus défavorables, jusqu'à 151,150× sur CPU et 1241,859× sur GPU dans les cas les plus favorables. Ce travail ouvre de nouvelles perspectives pour l'application d'architectures IA complexes dans des contextes industriels, tout en offrant une méthodologie accessible et reproductible pour les entreprises souhaitant adopter des solutions d'IA avancées.

Mots clefs : Apprentissage profond, Apprentissage profond distribué, Calcul haute performance, Industrie 4.0, Vision par ordinateur