Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Réentraînement conscient de la quantification : un arbitrage entre pré-entraînement et affinage des modèles de langue spécialisés

Xavier Pillet[1,3] Cédric Gernigon[1] Anastasia Volkova[2]
Richard Dufour[1] Adeline Granet[3] Nicolas Greffard[3]
(1) LS2N - Nantes Université, Nantes, France
(2) Inria, Laboratoire CITI, INSA Lyon, Villeurbanne, France
(3) Valeuriad, Nantes, France
nom.prénom@ls2n.fr, nom.prénom@inria.fr, prénom.nom@valeuriad.fr

RÉSUMÉ

La quantification est une technique largement adoptée pour réduire l'empreinte mémoire et le coût computationnel des réseaux de neurones. Si la quantification de modèles pré-entraînés s'avère efficace, un réentraînement est souvent nécessaire pour les formats de quantification extrême. L'affinage (*fine-tuning*), quant à lui, permet d'adapter des modèles généralistes à des domaines spécifiques, bien que la quantification puisse dégrader considérablement leurs performances. Ce travail étudie le coût d'entraînement des modèles de langue ajustés et quantifiés. La formalisation du compromis calculatoire entre l'adaptation au domaine et l'affinage, permet de démontrer que les points de contrôle spécialisés (checkpoints) présentent une plus grande robustesse au bruit de quantification. Ces résultats établissent un schéma directeur viable pour le déploiement de modèles de TAL biomédicaux performants dans des environnements embarqués aux ressources limitées.

ABSTRACT

Quantization-aware training : a tradeoff between training and fine-tuning for domain-specific language models

Here the title in English. Quantization is a widely adopted technique to reduce memory footprint and computational cost in neural networks. While quantizing pre-trained models is effective, retraining is often required for extreme quantization formats. Fine-tuning, on the other hand, enables the adaptation of general-purpose models to specific domains, but quantization can significantly degrade their performance. In this work, we investigate the training cost of fine-tuned and quantized language models. By formalizing the computational trade-off between domain adaptation and fine-tuning, we demonstrate that domain-specialized checkpoints exhibit greater robustness to quantization noise. Our findings establish a viable blueprint for deploying high-performance biomedical NLP models in resource-constrained, edge environments.

MOTS-CLÉS : Quantification, QAT, TAL Biomédical, Modèles de type BERT.

KEYWORDS: Quantization, QAT, Biomedical NLP, BERT-based model.

ARTICLE ACCEPTÉ À : AccML 2026.
URL : https://github.com/XavierValeuriad/Quantization-Aware-Pre-Training

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 41–41.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.