Détail de l'offre : Stage de Fin d'Études en Optimisation de l'IA pour l'Obtention d'Un Master 2 ou Diplôme d'Ingénieur H/F

Recruteur CEA
Ville , Bouches-du-Rhône
Référence 2024-34106
Titre de l'offre Stage de Fin d'Études en Optimisation de l'IA pour l'Obtention d'Un Master 2 ou Diplôme d'Ingénieur H/F
Description de la mission Le CEA est un acteur majeur de la recherche, au service des citoyens, de l'économie et de l'Etat.

Il apporte des solutions concrètes à leurs besoins dans quatre domaines principaux : transition énergétique, transition numérique, technologies pour la médecine du futur, défense et sécurité sur un socle de recherche fondamentale. Le CEA s'engage depuis plus de 75 ans au service de la souveraineté scientifique, technologique et industrielle de la France et de l'Europe pour un présent et un avenir mieux maîtrisés et plus sûrs.

Implanté au coeur des territoires équipés de très grandes infrastructures de recherche, le CEA dispose d'un large éventail de partenaires académiques et industriels en France, en Europe et à l'international.

Les 20 000 collaboratrices et collaborateurs du CEA partagent trois valeurs fondamentales :

- La conscience des responsabilités
- La coopération
- La curiosité

This internship proposes to explore a dual approach to optimizing ViTs by combining two complementary techniques : Token Pruning and Mixed Precision. Token pruning aims to reduce the amount of information processed at each layer by dynamically removing redundant or irrelevant tokens, thereby alleviating the computational load without significantly compromising performance. At the same time, mixed precision lets you use lower-precision number formats (like going from 32-bit precision to 16-bit or 8-bit) to save memory and speed up computations. This is possible while still keeping enough accuracy for vision tasks.

The goal of this internship is to design, implement, and evaluate the effectiveness of a dual approach within a vision transformer model to achieve an optimal balance between computational efficiency and predictive performance. The laboratory, which has experience working with quantified Vits models, has already developed a token reduction approach that has shown promising results for semantic segmentation tasks. The adaptation of state-of-the-art solutions will BE applied at different levels : at the encoder level, by integrating mixed precision quantization of operators, and at the decoder level, by adapting the model head to the quantized encoder to ensure consistency in information processing. Finally, benchmarking tests (FPS, mIOU, Params, MACC, FLOPS) will BE conducted on an embedded NVIDIA Orin card to evaluate the generalization capabilities of the token reduction model.

In this context, the objectives of the internship are :
A survey of the techniques for token reduction
A survey of the techniques for mixed precision quantification;
Benchmarking tests (FPS, mIOU, Params, MACC, FLOPS) of models with selected optimization techniques;
Develop a new frugal approach that challenges the state-of-the-art (SoTA);
Implementation on embedded chip type NVIDIA Jetson Orin.

#Token #TokenPruning
#MixedPrecision
#VIT #VisionTransformers #EfficientVisionTransformers
#ModelOptimization
#DeepLearning
#NeuralNetworks
#AIOptimization

	#MachineLearning
	#ModelCompression
	#ReducedComplexity
	#EnhancedPerformance
	Requested profile : Master degree (Bac +5)Pour postuler cliquer ici.
Type de contrat	Stage
Télétravail	Non spécifié
Profil	Ingénieur(e) CAO/DAO
Localisation	91120, PALAISEAU
Pays	France
Expérience	Expérimenté (3-10 ans)
Profil	Ingénieur(e) CAO/DAO
Fonction	Ingénieur(e) CAO/DAO
Secteur	Industries autres