

Deep Learning Optimisé - Jean Zay

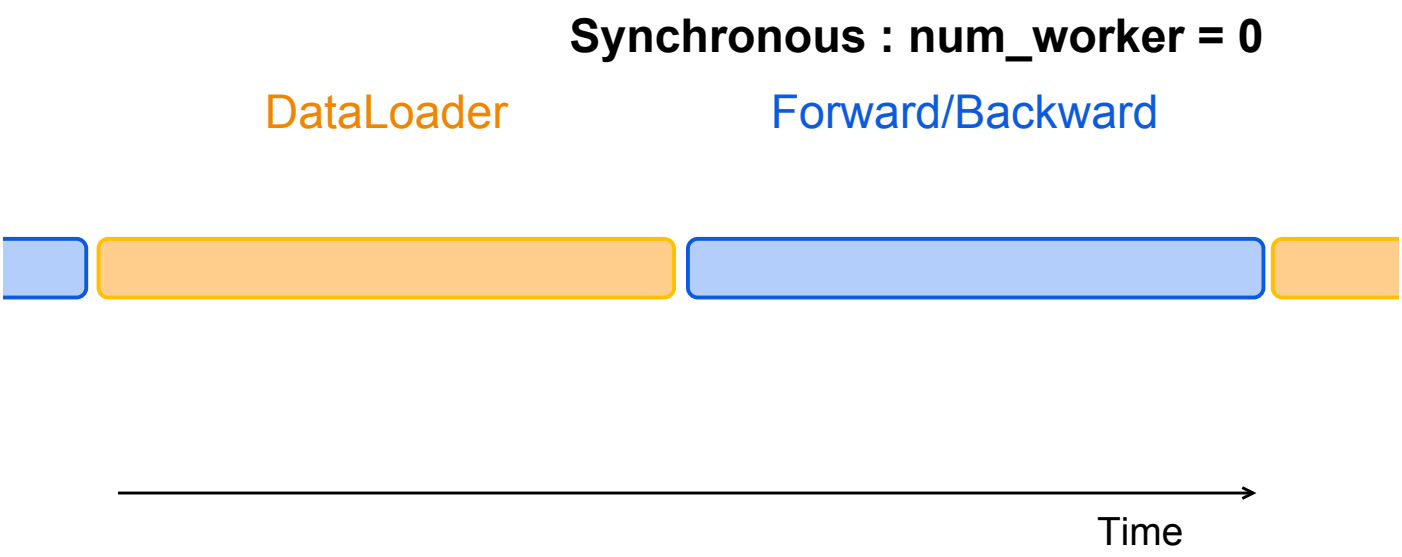
Conclusion



INSTITUT DU
DÉVELOPPEMENT ET DES
RESSOURCES EN
INFORMATIQUE
SCIENTIFIQUE



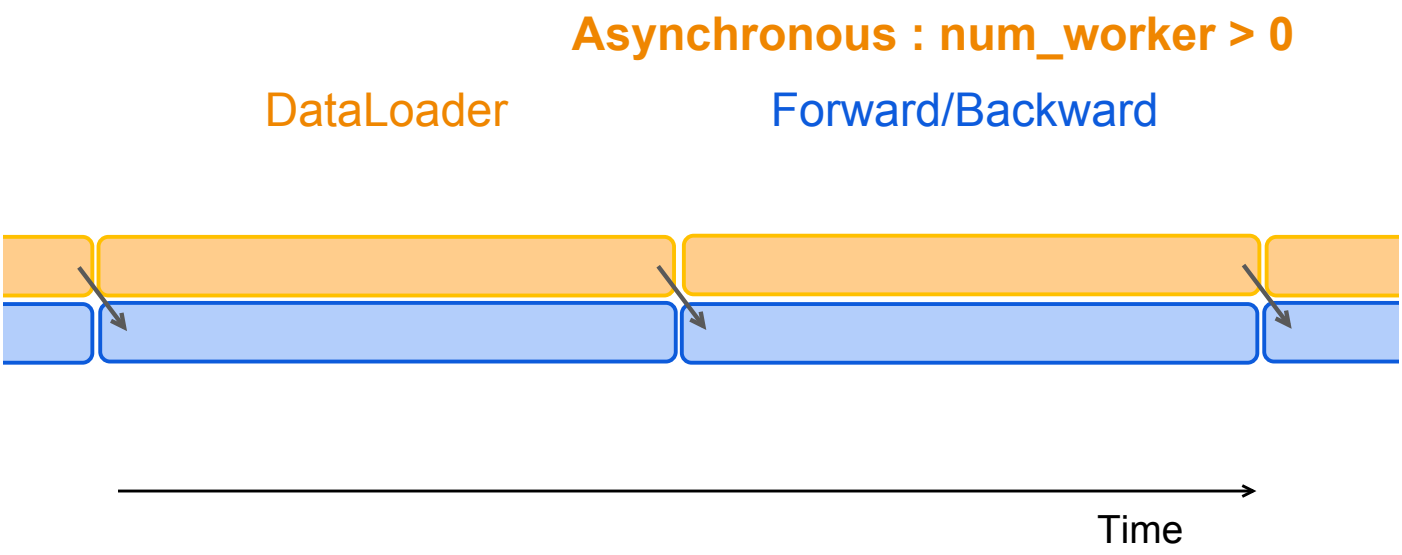
Conclusion



Bien optimiser et équilibrer la charge de travail du CPU et du GPU. Limiter les temps d’inactivité de chacun

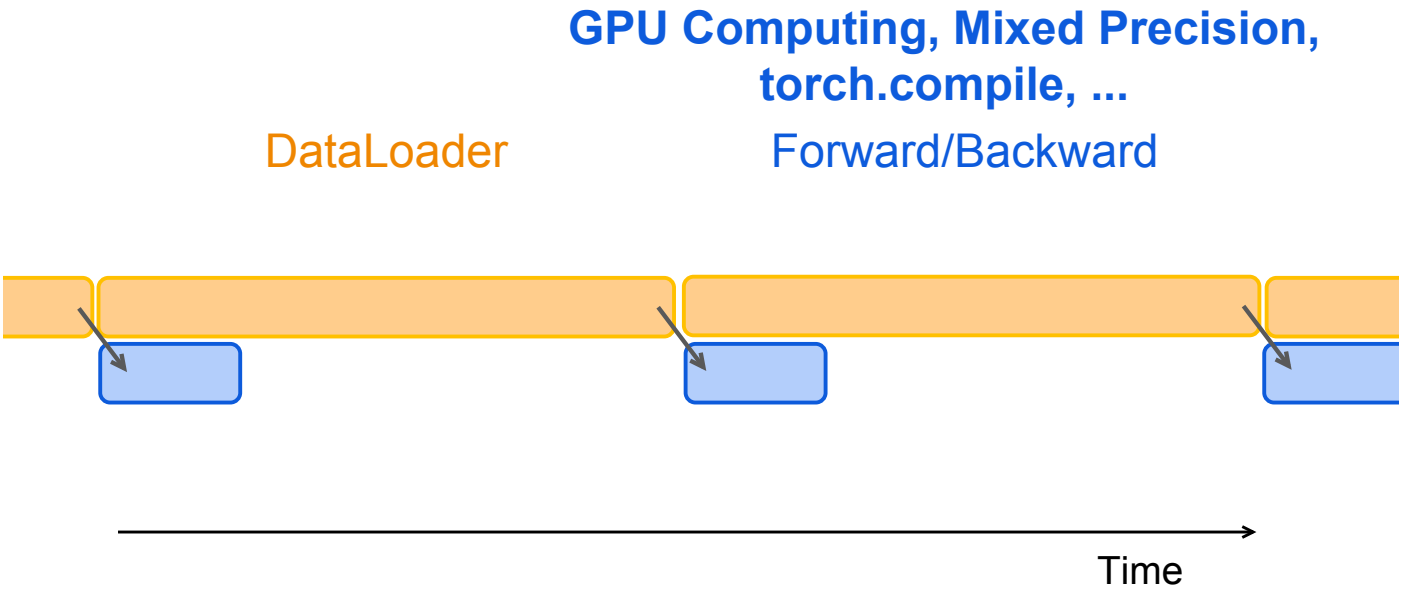
Importance de la visu et du profiling pour identifier les goulets d’étranglements

Conclusion



Asynchronisme: batch prêt quand vous en avez besoin

Conclusion



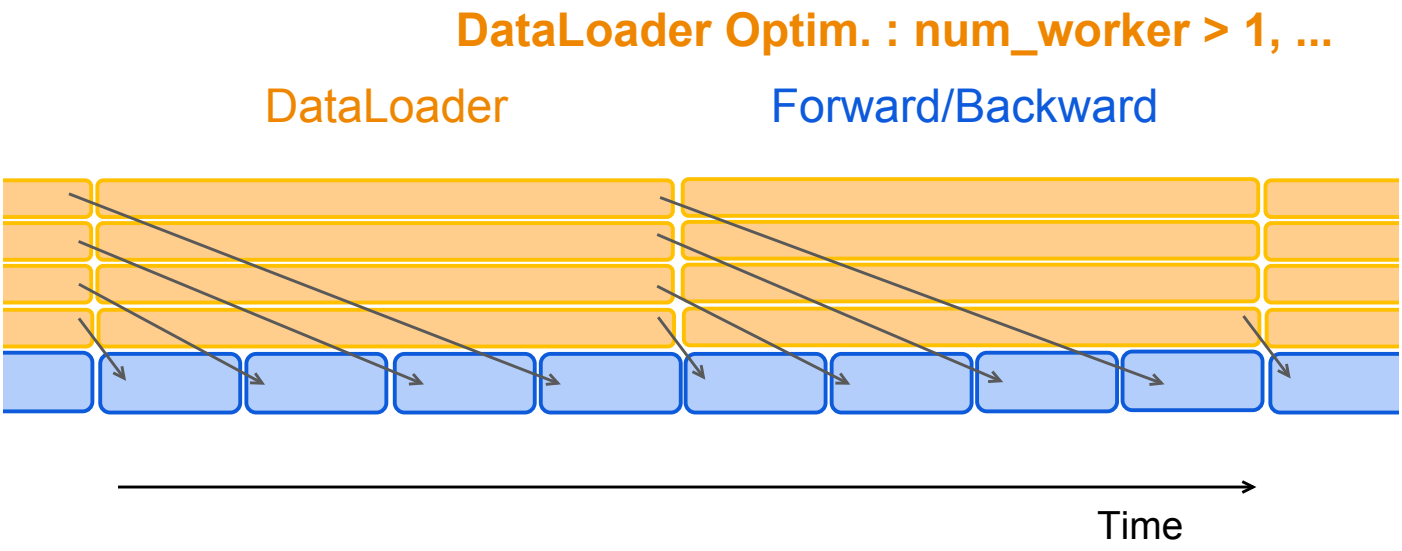
Différentes techniques pour accélérer les calculs:

Accélération GPU.

Compilation à la volée (tracer graph de calcul et le compiler/optimiser en language machine plus rapide que de l'interprété)

Précision mixte: Changer sur combien d'octets vous encodez vos poids pour gagner du temps de calcul et de la mémoire sans trop dégrader les perfs

Conclusion



Plus de workers = plus de batch prêts

Couplable à du prefetch pour préparer des batchs en avance au cas où le CPU viendrait à ralentir plus tard.

Conclusion

Training take too long !!!



Increase your batch size

Plus grands batch pour des pas plus grands (voit plus de données d’un coup)

Conclusion

Training take too long !!!



Increase your batch size

CUDA Out Of Memory !!!



Decrease your batch size

Batch trop grand = réduire taille de batch pour éviter OOM

Conclusion



For Small Model !!!

10s or 100s M Params

Distributed Data Parallelism

Pour de petits modèles à l’empreinte mémoire GPU faible: DDP

Conclusion



For Small Model !!!

10s or 100s M Params

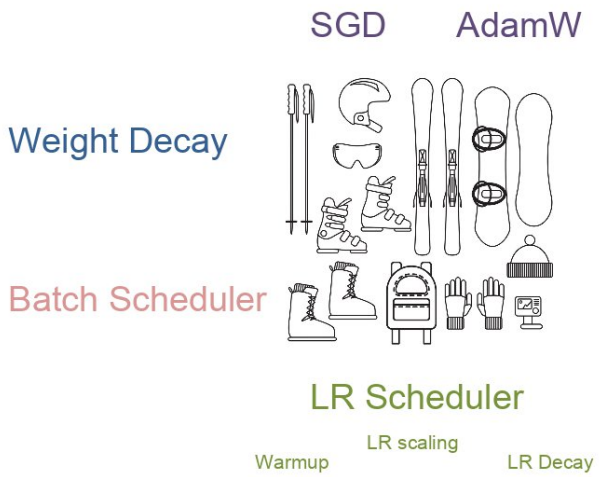
Distributed Data Parallelism



⟹ Large Batch !!

DDP = plus gros batchs, mais attention de larges batchs posent problème

Conclusion





Large batch = moyennes de gradients proches de 0

==> Rester bloqués dans des Sharp Minima

==> Utiliser des optimisers adaptés pour les large batchs (LAMB)

Conclusion



For Large Model !!!

> 1G Params

ZeRO

FSDP

Model Parallelisms

Pipeline Parallelism

Tensor Parallelism

Pour les gros modèles à l’empreinte mémoire élevéé:

Distribuer le modèle avec de la FSDP ou du Parallélisme de modèle