onprem, pour faire tourner facilement des LLM

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 20 septembre 2023

https://www.bortzmeyer.org/onprem-debut.html

Vous le savez, accéder à un LLM en ligne soulève de nombreux problèmes de dépendance et de confidentialité. La solution est de le faire tourner en local. Mais les obstacles sont nombreux.

Le premier obstacle est évidemment commercialo-juridique. Tous les modèles ne sont pas récupérables localement. Ainsi, les modèles derrière ChatGPT ou Midjourney ne semblent pas accessibles, et on doit les utiliser en ligne, confiant ses questions (et donc les sujets sur lesquels on travaille) à une entreprise commerciale lointaine. Même quand ils sont téléchargeables, les modèles peuvent avoir des restrictions d'utilisation (plusieurs entreprises affirment que leur modèle est "*open source*", terme qui a toujours été du baratin marketing et c'est de toute façon une question compliquée pour un LLM <https://framablog.org/2023/07/31/que-veut-dire-libre-ou-open-source-pour-un-grand-modele-de-langa>). Mais il existe plusieurs modèles téléchargeables et utilisables sous une licence plus accessible, ce qui nous amène au deuxième obstacle.

L'IA n'est pas décroissante ! Même si la consommation de ressources est bien plus faible pour faire tourner un modèle que pour l'entrainer, elle reste élevée et vous ne ferez pas tourner un LLM sur votre Raspberry Pi. La plupart du temps, 32 Go de mémoire, un processeur multi-cœurs rapide et, souvent un GPU Nvidia seront nécessaires. Et si vous ne les avez pas ? Des choses pourront tourner mais pas forcément de manière optimale. (Les exemples qui suivent ont été faits sur un PC/Ubuntu de l'année, avec 32 Go de mémoire, 8 cœurs à 3 Ghz, et un GPU Intel, apparemment pas utilisé ; il faut que je creuse cette question.)

Et le troisième obstacle est la difficulté d'utilisation des logiciels qui font tourner un modèle et, en réponse à une question, produisent un résultat. C'est peut-être l'obstacle le moins grave, car les progrès sont rapides. Quand on lit les articles où des pionniers faisaient tourner un LLM sur leur PC il y a un an, et qu'on voit à quel point c'était compliqué, on se réjouit des progrès récents. Ainsi, le premier septembre, a été publiée la première version de onprem <https://pypi.org/project/onprem/>, un outil libre pour faire tourner relativement facilement certains LLM. onprem est conçu pour les développeurs Python, et il faudra donc écrire quelques petits scripts (mais d'autres outils, comme open-interpreter <https://github.com/KillianLucas/open-interpreter>, ne nécessitent pas de programmer).

Démonstration. On installe onprem :