CroissantLLM : Des chercheurs de CentraleSupélec lancent un modèle d'IA open source et bilingue

Baptisé CroissantLLM, le modèle de langage développé par des chercheurs de CentraleSupélec apporte une alternative aux modèles poussés par les géants de la tech américains. La particularité de ce LLM : être entraîné sur autant de contenus en anglais qu'en français. Les chercheurs ont notamment bénéficié de la puissance de calcul du supercalculateur Jean Zay.

Célia Séramour
04 mars 2024 \ 17h45
3 min. de lecture

© MICS - L'équipe de recherche du laboratoire MICS de CentraleSupélec a développé avec Illuin Technology un modèle de langage (LLM) appelé CroissantLLM disponible sur la plateforme Hugging Face.

Les équipes de recherche du laboratoire MICS de CentraleSupélec ont développé conjointement avec plusieurs partenaires académiques un grand modèle de langage (LLM) appelé CroissantLLM. Disponible sur la plateforme Hugging Face, ce modèle se présente comme souverain et open source. Il a ainsi été développé par des Français et entraîné sur le supercalculateur Jean Zay.

Par ailleurs, les jeux de données sont français et publics, ce qui en fait réellement un modèle ouvert, contrairement par exemple à Llama 2 ou aux modèles de Mistral AI. Ces data sets proviennent de données juridiques, administratives, culturelles, commerciales, scientifiques et de traduction, précise Manuel Faysse, qui a participé au développement de ce LLM.

Un modèle bilingue français-anglais performant

Pré-entraîné sur un ensemble de 3000 milliards de tokens anglais et français, CroissantLLM compte 1,3 milliard de paramètres, bien loin des 175 milliards de paramètres de la version GPT-3.5 d'OpenAI. Précisons qu'il a été entraîné sur autant de contenus en français que de contenus en anglais, ce qui lui permet d'intégrer et de maîtriser les spécificités de la langue et de la culture françaises.

Basé sur une architecture de type Llama, le modèle est finalement plus petit que ceux publiés ces derniers mois. Et ce choix a été fait par les chercheurs pour une bonne raison : pousser à une meilleure adoption du modèle grâce à un fonctionnement sur du matériel grand public. *"Si l'on regarde les téléchargements de HuggingFace, les modèles les plus téléchargés ne sont pas les plus performants (Llama2-70B, Mixtral 8x7B) mais plutôt les plus petits (Llama2-7B, Mistral 7B) qui sont plus faciles et moins coûteux à servir et à ajuster",* constate Manuel Faysse.

Capable de fonctionner sur des CPU et des appareils mobiles

Les chercheurs ont donc fait le pari de proposer un modèle avec "peu" de paramètres et capable de fonctionner rapidement sur des serveurs GPU bas de gamme, tout en gardant un débit élevé et une faible latence. CroissantLLM peut également fonctionner sur des CPU ou même des appareils mobiles avec des vitesses décentes, indiquent les chercheurs, ce qui en fait un modèle économe en énergie.

Bien évidemment, il ne faut pas s'attendre à des capacités de raisonnement, de mathématiques et de code qui soient égales à d'autres modèles beaucoup plus grands. L'équipe de chercheurs estime qu'*"il sera parfait pour des applications industrielles plus spécifiques, des traductions ou même des capacités de chat dans lesquelles les gros canons ne sont pas toujours demandés".*

Un benchmark créé pour l'évaluation des performances du modèle en français

Pour évaluer la performance du modèle en français, les chercheurs ont également lancé un benchmark d'évaluation dédié et baptisé FrenchBench. Il est composé d'un ensemble de tâches de classification et de génération et couvre divers aspects de la performance des modèles en langue française. Sur la section à choix multiples de FrenchBench – qui se concentre sur le raisonnement, les connaissances factuelles et les capacités linguistiques – CroissantLLM atteint ainsi de meilleures performances que d'autres modèles de taille similaire.

Toujours dans cet objectif de transparence, les chercheurs ont publié des bases de code et des dizaines de points de contrôle pour différentes tailles de modèles, distributions de données d'entraînement et étapes d'entraînement, ainsi que des modèles de Chat affinés. *"Nous évaluons notre modèle à l'aide du framework FMTI et validons 81 % des critères de transparence",* précisent les chercheurs.

Les prémices d'autres recherches sur le développement de modèles bilingues

In fine, CroissantLLM et les artefacts associés visent également à être un support pour favoriser la poursuite des recherches sur les modèles de langage multilingues ainsi que la compréhension de l'impact des données de pré-entraînement sur les connaissances internes. En attendant, les entreprises peuvent accéder à deux versions de CroissantLLM – la version de base et une version fine-tuned pour un chatbot – depuis la plateforme Hugging Face.