

Des IA plus efficaces et parcimonieuses

3.9.2026 - | École Polytechnique

Vicky Kalogeiton, chercheuse au Laboratoire d'informatique de l'École polytechnique (LIX*), vient de recevoir une bourse Starting Grant du Conseil européen de la recherche (ERC). Le but de son projet de recherche : rendre les IA génératives d'images plus efficaces, plus accessibles et plus durables.

« Plus il y a de données, plus le modèle est gros, meilleur sont les résultats. » Voici, à grand trait, le mantra qui guide l'évolution des modèles d'intelligence artificielle depuis plusieurs années. L'irruption puis l'amélioration des IA comme ChatGPT, Claude ou Mistral, entraînées sur de très grands ensembles de données et utilisant de substantielles ressources de calcul, illustre ce paradigme. Mais il laisse une question sans réponse scientifique : peut-on construire des modèles permettant d'obtenir des résultats de haute qualité et contrôlables en empruntant une voie plus efficace ?

De la taille à l'efficacité des modèles

Dans un monde aux ressources finies où la durabilité constitue un enjeu pour l'avenir, cette question revêt beaucoup d'importance. C'est la motivation du projet de recherche de Vicky Kalogeiton, financé par le Conseil européen de la recherche. Son nom : FLASH (From scaling to efficiency laws for visual synthesis). Il s'agit de passer du paradigme de la taille à celui de l'efficacité pour la génération visuelle (images, vidéos, etc...).

« Il y a trois axes qu'il s'agit d'optimiser. Les données, qui servent à entraîner le modèle, le modèle lui-même qui calcule, et enfin les interactions, c'est-à-dire le jeu de ping-pong entre l'utilisateur et l'IA pour arriver au résultat souhaité » explique la chercheuse du LIX.

Dans un travail préliminaire, Vicky Kalogeiton et ses collègues ont montré qu'il était possible d'utiliser 1000 fois moins de données d'entraînement tout en obtenant de très bonnes performances (1). Comment diminuer cette taille en gardant les mêmes performances ? A quel point faut-il des images diverses ? Quel est l'impact d'avoir des images dupliquées ou du type d'image (des paysages naturels ou au contraire des figures abstraites).

Guider par les données

Ensuite, il s'agit d'optimiser le modèle en le « guidant » par les données. Si les images d'entraînement sont généralement annotées (par exemple, « une chaîne de montagnes enneigées »), les méthodes d'entraînement actuelles ne tiennent pas suffisamment compte de la structure et des caractéristiques des données elles-mêmes, pouvant ainsi entraîner une utilisation inefficace des ressources. « L'objectif sera d'extraire des informations supplémentaires des données afin de mieux guider l'entraînement du modèle » souligne Vicky Kalogeiton.

Pour un utilisateur, le résultat final est souvent obtenu après de multiples « prompts », ou interactions, avec l'IA. Comment rendre ces échanges moins longs tout en facilitant l'accessibilité ? L'IA pourrait demander des clarifications lorsque c'est nécessaire avant de générer une image, ou encore fournir d'abord une esquisse.

Des applications en robotique

En plus des applications de génération d'images ou de vidéos, le projet FLASH explorera également les applications de l'IA dans le domaine de la robotique, où les données, les capacités de calcul et les interactions sont particulièrement limitées. Par exemple, à partir d'images de l'environnement d'un bras robotique, un modèle IA doit pouvoir comprendre ce qui l'entoure et fournir la suite d'actions nécessaires pour que le bras saisisse un objet.

Enfin, FLASH ouvre la voie à des modèles embarqués plus efficaces, capables de fonctionner directement sur des appareils mobiles, ce qui permet de réduire la dépendance vis-à-vis des serveurs externes tout en renforçant la confidentialité et en diminuant la consommation d'énergie.

(1) Lucas Degeorge, Arijit Ghosh, Nicolas Dufour, David Picard, and Vicky Kalogeiton. How far can we go with imagenet for text-to-image generation ? In arXiv preprint arXiv:2502.21318 , 2025

*LIX : une unité mixte de recherche CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France

<https://www.polytechnique.edu/actualites/des-ia-plus-efficaces-et-parcimonieuses>