Interaction retardée dans l'encodeur du Transformer pour répondre efficacement aux questions dans un domaine ouvert

Wissam Siblini*, Mohamed Challal*
Charlotte Pasqual*

*Worldline Lyon
wissam.siblini@worldline.com

Résumé. La tâche de question-réponse sur un large corpus de documents (par exemple Wikipedia) est un défi majeur en informatique. Bien que les modèles de langage basés sur le Transformer tels que Bert aient montré une capacité à surpasser les humains pour extraire des réponses dans des petits passages de texte pré-sélectionnés, ils souffrent de leur grande complexité si l'espace de recherche est beaucoup plus grand. La façon la plus répandue de faire face à ce problème consiste à ajouter une étape préliminaire de recherche d'information pour filtrer fortement le corpus et ne conserver que les passages pertinents. Dans cet article, nous proposons une solution plus directe et complémentaire qui consiste à modifier l'architecture des modèles à base de Transformer pour permettre une gestion plus efficace des calculs. Les modèles qui en résultent sont compétitifs avec ceux d'origine et permettent, en domaine ouvert, une accélération significative des prédictions et parfois même une amélioration de la qualité de réponse.

1 Introduction

Depuis la parution du papier *Attention is all you need* de Vaswani et al. (2017), le modèle du Transformer n'a cessé de gagner en popularité dans le domaine du traitement automatique du langage naturel (TALN). Il est entraîné en deux étapes, une première de pré-entraînement coûteuse en temps et en données permettant de construire une représentation contextualisée des mots d'un vocabulaire, puis une seconde d'affinage moins onéreuse qui le spécialise dans une tâche de TALN bien précise. C'est de cette façon que Bert (Devlin et al., 2019), partie encodeur de l'architecture du Transformer, a montré une performance impressionnante sur de nombreuses tâches difficiles de compréhension du langage et a séduit un très large public par sa polyvalence et sa facilité d'utilisation. Des récentes variantes comme Albert (Lan et al., 2019) vont jusqu'à produire des réponses de meilleure qualité que l'humain.

Nous nous intéressons dans ce papier à la tâche de recherche d'une réponse à une question utilisateur dans un grand ensemble de documents textuels (par exemple, dans les millions de pages de l'encyclopédie Wikipedia). Les modèles de langage comme Bert sont efficaces sur une sous-tâche appelée Extractive Question Answering (eQA) dont le jeu de données de référence est SQuAD (Rajpurkar et al., 2016) : étant donnée une paire question-document, le but est de localiser la réponse dans le document. Mais sur notre tâche cible, appelée Open Domain