Attention Chaînée et Causale pour un Suivi Efficace des Entités

Erwan Fagnou[1] Paul Caillon[1] Blaise Delattre[1,2] Alexandre Allauzen[1,3]
(1) LAMSADE, Université Paris Dauphine - PSL
(2) Foxstream, Vaulx-en-Velin, France
(3) ESPCI PSL, Paris, France
prénom.nom@dauphine.psl.eu

RÉSUMÉ

Ce travail met en évidence une limitation théorique des transformers pour les tâches de suivi d'entités, montrant qu'ils nécessitent $\log_2(n+1)$ couches pour gérer n changements d'état. Pour surmonter cette contrainte, nous proposons ChaCAL (Chain and Causal Attention Layer), une modification de l'attention standard qui l'interprète comme une matrice d'adjacence, permettant de capturer efficacement les dépendances longues avec une seule couche. Les expériences menées sur un jeu de données synthétique et un autre de suivi d'objets démontrent que ChaCAL surpasse les transformers classiques en réduisant la profondeur du modèle, tout en maintenant des performances compétitives sur une tâche de modélisation du langage. Cette approche optimise l'efficacité des modèles tout en réduisant leur coût computationnel.

ABSTRACT

Chain and Causal Attention for Efficient Entity Tracking

This paper investigates the limitations of transformers for entity-tracking tasks in large language models. We identify a theoretical constraint, showing that transformers require at least $\log_2(n+1)$ layers to handle entity tracking with n state changes. To address this issue, we propose an efficient and frugal enhancement to the standard attention mechanism, enabling it to manage long-term dependencies more efficiently. By considering attention as an adjacency matrix, our model can track entity states with a single layer. Empirical results demonstrate significant improvements in entity tracking datasets while keeping competitive performance on standard natural language modeling. Our modified attention allows us to achieve the same performance with drastically fewer layers. Additionally, our enhanced mechanism reveals structured internal representations of attention. Extensive experiments on both toy and complex datasets validate our approach. Our contributions include theoretical insights, an improved attention mechanism, and empirical validation.

MOTS-CLÉS : Suivi d'entités, Transformers, Mécanisme d'attention efficace, Efficacité computationnelle, Frugalité.

KEYWORDS: Entity Tracking, Transformers, Efficient attention mechanism, Computational efficiency, Frugality.

ARTICLE : **Accepté à** EMNLP 2024.

Lien vers l'article original: https://aclanthology.org/2024.emnlp-main.731/