Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

LOCOST: Modèles Espace-État pour le Résumé Abstractif de Documents Longs

accepté à EACL 2024

Florian Le Bronnec[1, 2] Song Duong[1, 6] Alexandre Allauzen[2]
Vincent Guigue[5] Alberto Lumbreras[6] Laure Soulier[1] Patrick Gallinari[1, 6]
(1) Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
(2) Miles Team, Université Paris-Dauphine, Université PSL, CNRS, LAMSADE, 75016 Paris, France
(5) AgroParisTech, UMR MIA-PS, Palaiseau, France
(6) Criteo AI Lab, Paris, France
florian.le-bronnec@dauphine.psl.eu, s.duong@criteo.com

RÉSUMÉ

Les modèles espace-état constituent une alternative peu coûteuse en termes de complexité de calcul aux transformeurs pour le codage de longues séquences et la capture de longues dépendances. Nous proposons LOCOST : une architecture encodeur-décodeur basée sur des modèles espace-état pour la génération de textes conditionnels avec de longues entrées contextuelles. Avec une complexité de calcul de $\mathcal{O}(L \log L)$, cette architecture peut traiter des séquences beaucoup plus longues que les modèles de référence qui sont basés sur des modèles d'attention parcimonieux. Nous évaluons notre modèle sur une série de tâches de résumé abstractif de longs documents. Le modèle atteint un niveau de performance qui est 93-96% comparable aux transformeurs parcimonieux les plus performants de la même taille tout en économisant jusqu'à 50% de mémoire pendant l'apprentissage et jusqu'à 87% pendant l'inférence. En outre, LOCOST traite efficacement les entrées dépassant *600K* tokens au moment de l'inférence, établissant de nouveaux résultats de référence sur le résumé de livre complet et ouvrant de nouvelles perspectives pour le traitement des entrées longues.

ABSTRACT

LOCOST : State-Space Models for Long Document Abstractive Summarization.

State-space models are a low-complexity alternative to transformers for encoding long sequences and capturing long-term dependencies. We propose LOCOST : an encoder-decoder architecture based on state-space models for conditional text generation with long context inputs. With a computational complexity of $\mathcal{O}(L \log L)$, this architecture can handle significantly longer sequences than state-of-the-art models that are based on sparse attention patterns. We evaluate our model on a series of long document abstractive summarization tasks. The model reaches a performance level that is 93-96% comparable to the top-performing sparse transformers of the same size while saving up to 50% memory during training and up to 87% during inference. Additionally, LOCOST effectively handles inputs exceeding *600K* tokens at inference time, setting new state-of-the-art results on full-book summarization and opening new perspectives for long input processing.

MOTS-CLÉS : modèles espace-état, résumé abstractif de documents longs.

KEYWORDS: state-space models, long document abstractive summarization.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 11–11.