Représentations Hiérarchiques pour les longs documents

Iskandar Boucharenc[1]
(1) LISN, Campus Universitaire bâtiment 507, Rue du Belvédère, 91400 Orsay, France
iskandar.boucharenc@lisn.fr

RÉSUMÉ

Les progrès récents en recherche d'information (RI) ont largement bénéficié des représentations vectorielles denses issues de systèmes du traitement automatique des langues. Les systèmes basés sur l'architecture transformeur excellent dans la compréhension de textes à l'échelle des jetons. Cependant, l'extension à des unités textuelles de niveaux supérieurs (*e.g.* fragments, phrases, documents ...) reste coûteuse en termes de mémoire et de calcul. Dans cet article, nous étudions la pertinence des solutions existantes pour représenter les fragments. Puis, nous proposons une modification des tâches de pré-entraînement afin de capturer l'information d'ordre supérieur dans des jetons hiérarchiques spéciaux. À partir d'un texte découpé de manière hiérarchique, chaque niveau possède son propre vecteur de représentation, partagé avec ses sous-niveaux. Notre procédure de pré-entraînement permet aux représentations d'ordre supérieur d'apprendre la sémantique du fragment de texte à partir des niveaux inférieurs. La fonction de perte se base sur la divergence de Kullback–Leibler. L'affinage en aval de ces représentations hiérarchiques sur des tâches classiques de RI est simple et direct. La méthode proposée permet d'utiliser le même système pour toutes les étapes de la RI, tout en simplifiant les processus de classement et de récupération de passages. De plus, ces représentations devraient être suffisamment flexibles pour être utilisées dans des tâches à long contexte.

ABSTRACT

Hierarchical Prefixes for Long Document Representations

Recent advances in information retrieval have greatly benefited from dense representations of deep learning systems in natural language processing. Based on transformer architecture, modern systems are exceedingly good for token-level understanding. However, extending to higher-level text units remains prohibitively expensive in terms of memory and computation. In this paper, we study the strengths of existing solutions for sentence and chunk-level representations. We propose a slight modification of pre-training tasks to capture information in specialized hierarchical tokens. Given a text split following a hierarchical order, each level has its specific vector representation shared by its sublevels. The pre-training procedure allows higher-order representations to learn semantics from low-level semantics using Kullback–Leibler-based losses. Fine-tuning hierarchical vector representation on classical retrieval tasks is straightforward. Our methods allow the use of the same system and model for all stages in IR and simplify the ranking and retrieval process. Moreover, such representations are flexible enough for general long-context tasks.

MOTS-CLÉS : Représentation de document, Recherche d'information, Affinage frugal.

KEYWORDS: Document Representation, Information Retrieval, Parameter Efficient Fine-tuning.

ARTICLE ACCEPTÉ À : 47th European Conference on Information Retrieval (Doctoral Consortium).
URL : https://link.springer.com/chapter/10.1007/978-3-031-88720-8_28