LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders

Parishad BehnamGhader, Vaibhav Adlakha Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, Siva Reddy

Publié: 21 Août 2024

***Résumé*—Les modèles de langage à décodeur seul (LLM) sont performants en NLP, mais l'adoption pour les tâches d'incorporation de texte est lente. L'article propose LLM2Vec, une approche non supervisée pour transformer ces modèles en encodeurs puissants. L'approche comprend trois étapes : 1) L'attention bidirectionnelle, 2) la prédiction masquée du prochain jeton, 3) L'apprentissage contrastif. L'efficacité de LLM2Vec est démontrée sur des modèles de 1,3B à 8B paramètres, surpassant les encodeurs classiques et établissant une nouvelle performance sur le Massive Text Embeddings Benchmark (MTEB). Combiné avec l'apprentissage supervisé, LLM2Vec obtient des performances de pointe avec des données publiques, sans adaptation coûteuse.**

***Keywords*—LLM, Encodeur, Decodeur, Embeddings**

I. Introduction

LLM2VEC est une approche simple et non supervisée pour transformer les grands modèles de langage (LLMs) basés uniquement sur des décodeurs en encodeurs de texte puissants. Ils mettent en avant son efficacité en termes de données et de paramètres, ses performances empiriques solides sur diverses tâches, ainsi que sa capacité à atteindre des résultats à la pointe de l'état de l'art, aussi bien dans des contextes non supervisés que supervisés, sans nécessiter d'adaptations coûteuses ni de données synthétiques.

II. Contexte

L'article présente LLM2Vec, une méthode qui transforme des modèles de langage à décodeur seul (LLM) en encodeurs de texte efficaces. Les modèles spécifiquement mentionnés incluent

- S-LLaMA-1.3B,
- LLaMA-2-7B,
- Mistral-7B,
- et Meta-LLaMA-3-8B.

Cette approche non supervisée améliore la représentation contextuelle des textes, surpassant les modèles encodeurs traditionnels sur plusieurs tâches de traitement du langage naturel.

III. Problématiques

Le problème principal abordé par l'article est l'adoption limitée des grands modèles de langage (LLMs) basés uniquement sur des décodeurs pour les tâches d'encodage de texte, en raison de contraintes architecturales telles que l'attention causale, qui entrave la production de représentations contextuelles riches. L'étude propose de transformer ces modèles en encodeurs de texte efficaces grâce à l'approche LLM2Vec.

IV. Objectifs globale

L'objectif global de l'article est d'améliorer l'utilisation des modèles de langage à décodeur seul (LLM) dans le domaine du traitement du langage naturel, en les adaptant pour produire des représentations textuelles contextuelles de haute qualité. Cela vise à élargir les applications et l'efficacité des LLM, tout en surmontant les limitations de leur architecture initiale, afin de rendre ces modèles plus accessibles et performants dans des tâches variées.

V. Objectifs spécifiques

1) Transformer les LLM en encodeurs de texte : les auteurs proposent LLM2Vec comme une méthode efficace pour convertir des modèles de langage à décodeur seul en encodeurs capables de produire des représentations textuelles contextuelles de haute qualité.
2) Surpasser les modèles encodeurs traditionnels : Démontrer que les modèles transformés par LLM2Vec surpassent les modèles encodeurs classiques sur des tâches de traitement du langage naturel, notamment en termes de performance sur des benchmarks comme le Massive Text Embeddings Benchmark (MTEB).
3) Utiliser une approche non supervisée : Mettre en avant l'efficacité de LLM2Vec qui ne nécessite pas de données étiquetées, rendant la méthode accessible et applicable dans des contextes où les données annotées sont rares.
4) Analyser l'impact de LLM2Vec : Fournir une analyse approfondie des effets de LLM2Vec sur les représentations des modèles sous-jacents, y compris la capacité des LLM à capturer l'information des tokens futurs.
5) Promouvoir l'efficacité des LLM : Mettre en évidence que les modèles de langage à décodeur seul peuvent être adaptés de manière efficace et économique pour des tâches d'embedding de texte, tout en conservant leurs avantages en termes de taille et d'efficacité d'échantillonnage.

VI. Dataset

Le dataset de Massive Text Embedding Benchmark(MTEB) comprend 56 ensembles de données couvrant diverses tâches d'embedding textuel, telles que la récupération d'information, le clustering et la classification. Les catégories spécifiques incluent:

- Retrieval,
- Reranking,
- Clustering,

- Pair Classification,
- Classification,
- STS,
- et Summarization,

permettant une évaluation approfondie des performances de LLM2Vec.

VII. Solution proposée

1) Activation de l'attention bidirectionnelle : Permet aux tokens d'interagir avec toute la séquence et non seulement avec les tokens précédents.
2) Masked Next Token Prediction (MNTP) : Entraîne le modèle à prédire des tokens masqués, améliorant ainsi la contextualisation.
3) Unsupervised Contrastive Learning (SimCSE) : Optimise les embeddings en maximisant la similarité entre différentes vues d'un même texte.
   Ces trois étapes composent l'approche LLM2Vec, permettant d'adapter efficacement les LLMs decoder-only aux tâches d'embedding.

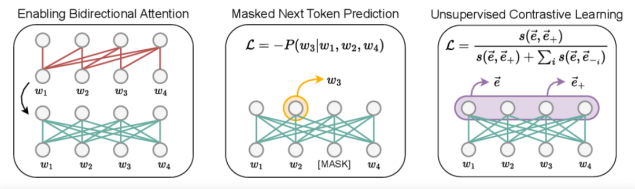

Fig. 1. The 3 steps of LLM2Vec

VIII. Implémentation

La méthode de LLM2Vec repose sur des modèles decoder-only pré-entraînés et utilise principalement PyTorch et la bibliothèque Hugging Face pour la manipulation des embeddings. Voici les modèles utilisés

TABLE I
Modèles utilisés

Modèles	Paramètres
S-LLaMA	1.3B
LLaMA-2	7B
Mistral	7B
Meta-LLaMA-3	8B

Source : Ces modèles sont récupérés depuis Hugging Face et adaptés avec LLM2Vec pour devenir des encodeurs de texte.

A. Activation de l'attention bidirectionnelle

- Par défaut, les modèles decoder-only utilisent une attention causale, empêchant chaque token de voir les suivants.
- Les auteurs remplacent le masque d'attention par une matrice d'attention bidirectionnelle (matrice de 1s), permettant à tous les tokens d'interagir entre eux.
- Cette modification est réalisée en manipulant directement les masques d'attention dans le modèle Hugging Face.

B. Entraînement avec la prédiction de token masqué (MNTP)

cette deuxième solution a comme objectif d'adapter le modèle pour utiliser l'attention bidirectionnelle efficacement, ainsi la méthode procède par:

- Masquer un pourcentage des tokens dans chaque séquence d'entrée.
- Remplacer le token masqué par un token spécial ([MASK]).
- Demander au modèle de prédire le token masqué en utilisant le contexte environnant.
- Données : L'entraînement est réalisé sur Wikitext-103.
- Optimisation : Fine-tuning avec LoRA (Low-Rank Adaptation) pour réduire le nombre de paramètres mis à jour.
- Framework : Implémenté avec PyTorch et Transformers de Hugging Face.

C. Apprentissage Contrastif Non Supervisé (SimCSE) dans LLM2Vec

L'apprentissage contrastif est utilisé pour améliorer la cohérence des représentations textuelles en apprenant à distinguer les phrases similaires des différentes. SimCSE (Simple Contrastive Sentence Embeddings) permet au modèle de produire des embeddings robustes sans supervision.

1) Principe du SimCSE: L'idée est d'exploiter une propriété simple :

- Si une phrase est encodée plusieurs fois avec un léger bruit (comme du dropout), ses embeddings doivent rester proches.
- Les embeddings d'autres phrases doivent être éloignés pour éviter la confusion entre concepts différents.

2) Méthode Appliquée dans LLM2Vec:

- Encodage Double avec Dropout
- Chaque phrase est passée deux fois à travers le modèle avec différents masques de dropout.
- Cela crée deux versions légèrement différentes de son embedding.

3) Optimisation par Perte Contrastive:

- La fonction de perte vise à maximiser la similarité entre les deux embeddings de la même phrase.
- Simultanément, elle minimise la similarité avec les embeddings des autres phrases du batch.
- Formule de la perte contrastive utilisée :

$$\mathcal{L} = -\log \frac{e^{s(\vec{e},\vec{e}_+)/\tau}}{\sum_i e^{s(\vec{e},\vec{e}_i)/\tau}} \quad (1)$$

où :

  - $s(\vec{e},\vec{e}_+)$ est la similarité entre la phrase et elle-même (après dropout).
  - $s(\vec{e},\vec{e}_-)$ est la similarité avec une autre phrase du batch (négatif).
  - τ est une température qui contrôle la séparation entre les classes.

4) Pooling pour Extraire l'Embedding Final:

- Après la transformation des tokens, un moyennage est appliqué pour produire un unique vecteur représentant la phrase.

IX. Configuration expérimentale

L'évaluation de LLM2Vec est réalisée sur le Massive Text Embedding Benchmark (MTEB), un benchmark couvrant 56 ensembles de données répartis en 7 catégories de tâches :

- Retrieval (récupération d'information)
- Reranking (réordonnancement)
- Clustering (regroupement)
- Pair classification (classification par paires)
- Classification
- STS (similarité sémantique entre textes)
- Summarization (résumé automatique)

X. Comparaison et métriques

- Comparaison avec Echo: une méthode concurrente.
- Métrique principale : Similarité cosinus. Méthodes de pooling analysées :
- EOS pooling
- Mean pooling
- Weighted mean pooling

LLM2Vec obtient les meilleurs résultats avec le mean pooling

XI. Résultats expérimentaux

1) L'activation de l'attention bidirectionnelle seule améliore déjà les performances de Mistral-7B et S-LLaMA-1.3B, sans besoin de réentraînement
2) L'intégration de la prédiction masquée de jetons (MNTP) améliore encore la qualité des représentations textuelles.
3) LLM2Vec surpasse les intégrations Echo, notamment après l'application des deux premières étapes (attention bidirectionnelle + MNTP).
4) Ces résultats confirment l'efficacité de LLM2Vec pour transformer les modèles LLM en encodeurs de texte performants.
5) SimCSE apporte un gain supplémentaire, notamment pour S-LLaMA-1.3B (+49,8 %), LLaMA-2-7B (+23,2 %) et Mistral-7B (+37,5 %) sur MTEB

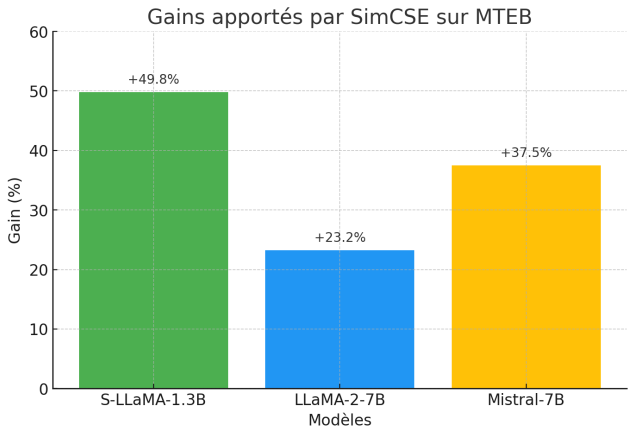

Fig. 2. Résultat de l'impact de Gain apportés par SimCSE sur MTEB

Ce graphique illustre les gains de performance obtenus grâce à l'application de la méthode SimCSE sur le benchmark MTEB pour trois modèles de langage différents. L'objectif est de visualiser les améliorations apportées par cette méthode pour chaque modèle.

Sur l'axe horizontal, on retrouve les trois modèles concernés : S-LLaMA-1.3B, LLaMA-2-7B et Mistral-7B. L'axe vertical indique les pourcentages de gain, avec une échelle allant de 0 % à 60 % afin de bien visualiser les écarts entre les performances des différents modèles.

Chaque barre colorée correspond à un modèle spécifique et représente le pourcentage d'amélioration obtenu grâce à SimCSE. On observe que le modèle S-LLaMA-1.3B, représenté en vert, enregistre le gain le plus élevé avec +49,8 %. Ce résultat démontre que l'application de SimCSE a considérablement amélioré les performances de ce modèle sur MTEB.

Le modèle Mistral-7B, en jaune, affiche un gain de +37,5 %, ce qui constitue également une amélioration significative. En revanche, le modèle LLaMA-2-7B, en bleu, présente un gain plus modeste de +23,2 %. Bien que cette augmentation soit moins marquée, elle reste tout de même notable, prouvant que SimCSE a un effet positif sur chacun des modèles.

Pour faciliter la lecture, chaque barre est annotée avec sa valeur exacte de gain, et une grille horizontale en pointillés permet de mieux estimer visuellement la hauteur des barres. Cette représentation permet ainsi de comparer efficacement l'impact de SimCSE sur les performances des différents modèles de langage évalués sur MTEB.

XII. Critiques Positives

- L'article propose une approche simple et non supervisée (LLM2Vec) pour transformer n'importe quel LLM uniquement décodeur en un encodeur de texte puissant.
- LLM2Vec se compose de trois étapes claires et intuitives : 1) activer l'attention bidirectionnelle, 2) la prédiction masquée du prochain jeton (MNTP), et 3) l'apprentissage contrastif non supervisé.
- La méthode ne nécessite aucune donnée étiquetée et se révèle très efficace en termes de données et de paramètres.
- LLM2Vec appliqué à des LLMs populaires (S-LLaMA-1.3B, LLaMA-2-7B, Mistral-7B, Meta-LLaMA-3-8B) surpasse largement les modèles à encodeur seul sur les tâches au niveau des mots (découpage, reconnaissance d'entités nommées, étiquetage des parties du discours).
- Sur le Massive Text Embeddings Benchmark (MTEB), les modèles transformés par LLM2Vec établissent un nouvel état de l'art pour les modèles non supervisés.
- En combinant LLM2Vec avec l'apprentissage contrastif supervisé, l'article obtient des performances de pointe sur MTEB parmi les modèles qui s'entraînent uniquement sur des données accessibles au public.
- L'article révèle une propriété intrigante de Mistral-7B, qui peut gérer l'attention bidirectionnelle sans aucun réglage fin.
- LLM2Vec se révèle plus performant que l'approche concurrente "Echo Embeddings" pour la plupart des modèles testés.
- L'application de LLM2Vec conduit à un entraînement plus efficace en termes d'échantillons lors du réglage fin supervisé.

- La méthode n'induit pas de surcharge de calcul supplémentaire au moment de l'inférence, contrairement à certaines approches concurrentes

XIII. Critiques Négatives

- L'article reconnaît que la lente adoption des LLMs uniquement décodeurs pour l'intégration de texte est en partie due à leur mécanisme d'attention causale, intrinsèquement limité pour produire des représentations contextualisées riches.
- L'activation naïve de l'attention bidirectionnelle sans entraînement nuit considérablement aux performances pour la plupart des modèles (à l'exception notable de Mistral-7B).
- Bien que la combinaison de MNTP et SimCSE donne de bons résultats pour les tâches au niveau des séquences, elle donne de moins bons résultats que la simple application de MNTP pour les tâches au niveau des mots.
- La grande taille des LLMs uniquement décodeurs peut entraîner une inférence considérablement plus lente.
- La grande dimension d'intégration de sortie de ces modèles (par exemple, 4096 pour Mistral-7B) les rend plus gourmands en mémoire et en calcul pour la création d'index vectoriels pour de grandes collections de documents.
- L'évaluation se concentre principalement sur l'anglais.
- Les détails complets des données de pré-entraînement des modèles (LLaMA-2-7B, Mistral-7B) n'étant pas accessibles au public, l'étendue de la contamination potentielle des données d'évaluation ne peut être entièrement déterminée.

XIV. Avantage Principal

L'avantage principal de LLM2Vec est sa capacité à transformer des LLMs uniquement décodeurs pré-entraînés en encodeurs de texte universels performants de manière simple, efficace en termes de paramètres, et sans nécessiter de données étiquetées coûteuses.

XV. Désavantage Principal

Le désavantage principal est le coût d'inférence potentiellement plus élevé et la consommation de mémoire plus importante associés à l'utilisation de grands modèles de langage pour l'intégration de texte, ainsi que la nécessité d'une adaptation spécifique (MNTP et SimCSE) pour exploiter pleinement le potentiel de ces modèles en tant qu'encodeurs bidirectionnels.

XVI. Optimisations Potentielles de l'approche LLM2Vec

Pour optimiser l'approche LLM2Vec, plusieurs solutions peuvent être envisagées comme l'étendre au multilingue via des corpus comme Wikipédia, réduire les coûts d'inférence par quantification ou distillation, et explorer des stratégies de pooling avancées. L'adaptation des données à des tâches ciblées, le renforcement de l'apprentissage supervisé et la gestion des contextes longs sont également clés. Enfin, l'usage de données synthétiques pour l'apprentissage contrastif peut améliorer la généralisation des représentations.

XVII. Conclusion

l'approche LLM2Vec offre une méthode prometteuse pour transformer des grands modèles de langage uniquement décodeurs en puissants encodeurs de texte grâce à des étapes clés d'adaptation. Afin de maximiser son potentiel et d'adresser certaines limitations, diverses pistes d'amélioration peuvent être explorées. Ces axes comprennent l'extension des capacités multilingues, l'optimisation de l'efficacité en termes de calcul et de stockage, la spécialisation des données d'adaptation pour des tâches spécifiques, l'étude de stratégies de pooling plus sophistiquées, une intégration accrue avec l'apprentissage supervisé et l'amélioration de la gestion des contextes longs. L'ensemble de ces investigations vise à rendre l'approche LLM2Vec encore plus robuste et performante pour la production d'embeddings de texte de haute qualité.

[1]
[2] [3]

References

[1] P. X. Yingfeng Luo,Tong Zheng, Z. T. Yongyu Mu, Bei Li, and D. B. Qinghong Zhang, Yongqi Gao, "Beyond decoder-only: Large language models can be good encoders for machine translation," in *Proc. Int. Conf. Industrial Instrumentation and Control (ICIC)*, 9 Mar 2025, p. 36. DOI: arXiv:2503.06594v1[cs.CL].

[2] N. M. Niklas Muennighoff1, Nouamane Tazi1, L. M. Loïc Magne1, and N. R. Nils Reimers2, "Mteb: Massive text embedding benchmark," in *Proc. Int. Conf. Industrial Instrumentation and Control (ICIC)*, 19 Mars 2023, p. 36. DOI: arXiv:2210.07316v3[cs.CL].

[3] L. N. Liang Wang, Nan Yang, X. L. Xiaolong Huang, Linjun Yang, and F. R. Linjun Yang, Rangan Majumder, Furu Wei, "Improving text embeddings with large language models," in *Proc. Int. Conf. Industrial Instrumentation and Control (ICIC)*, 31 May 2024, p. 20. DOI: arXiv:2401.00368v3[cs.CL].

XVIII. TRAVAUX CONNEXES

Titre	**Beyond Decoder-only: Large Language Models Can be Good Encoders for Machine Translation**
Auteurs et Année	Yingfeng Luo, Tong Zheng, Yongyu Mu, Bei Li, Qinghong Zhang, Yongqi Gao, Ziqiang Xu, Peinan Feng, Xiaoqian Liu, Tong Xiao, Jingbo Zhu (9 Mars 2025)
Méthodologies	Architecture hybride LaMaTE : LLM comme encodeur + décodeur NMT, adaptateur intermédiaire, entraînement en 2 étapes, benchmark ComMT pour évaluation multitache.
Résultats	LaMaTE surpasse NMT classiques et LLM fine-tunés, inférence 2.4×–6.5× plus rapide, 75% de mémoire KV cache économisée, forte généralisation sur divers types de traduction.
Titre	**MTEB: Massive Text Embedding Benchmark**
Auteurs et Année	Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers (19 Mars 2023)
Méthodologies	Création de MTEB avec 8 tâches, 58 datasets, 112 langues. Évaluation de 33 modèles (SimCSE, MPNet, GTR, ST5, etc.). Benchmark multilingue et multitâches, mesures variées selon tâche.
Résultats	Aucun modèle ne domine toutes les tâches. ST5-XXL domine en STS et classification. MPNet performant malgré sa petite taille. SGPT-msmarco meilleur en retrieval. LaBSE fort en bitext mining multilingue.
Titre	**Improving Text Embeddings with Large Language Models**
Auteurs et Année	Liang Wang et al. (31, Mai, 2024)
Méthodologies	Génération de données synthétiques via LLMs (GPT-4), fine-tuning de Mistral-7B avec perte contrastive, évaluation sur MTEB et BEIR.
Résultats	Nouveaux SOTA sur MTEB (+2%), performance compétitive avec données synthétiques uniquement, gestion de contextes longs (32k tokens), limite : coût d'inférence élevé.

TABLE II
ANALYSE SYNTHÉTIQUE DES TRAVAUX CONNEXES

Lien vers l'article