Transforming Decoder-LLM to Top-Tier Danish Encoder Models Using LLM2Vec Approach

Jesper Alkestrup, s133696
The Tech Collective

Introduction

The best-performing embedding models have shifted from being regular encoder-only models like BERT, to now exclusively being LLM-decoder models turned into encoders via various model tweaks and finetuning-steps.

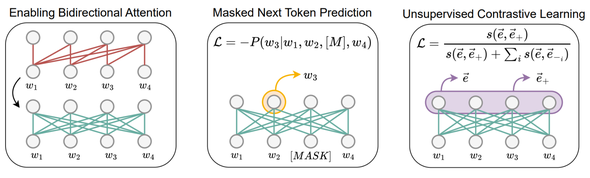
Overview of top-performing models on MTEB[1]

Embedding Model	Base Model	Pooling	Attention	Training Data Size	Score	Rank
NV-Embed-v2	Mistral-7B-v0.1	Trainable pooling layer	Bidirectional	1.1M (multiple runs)	72.31	1
BGE-en-icl (few shot query)	Mistral-7B-v0.1	EOS-Last token pool	Causal	1.5M (multiple runs)	71.67	2
gte-Qwen-1.5-7B-Instruct	Qwen2-7B	Not disclosed	Bidirectional	Not disclosed	70.24	3-10*
SFR-Embedding-2_R	Mistral-7B-v0.1	EOS-Last token pool	Causal	Not disclosed	70.31	5*
e5-mistral-7b-instruct	Mistral-7B-v0.1	EOS-Last token pool	Causal	1.8M (partly private)	66.63	26
LLM2Vec-Llama3-supervised	Llama-3-8B	Mean pool	Bidirectional	1.5M (public)	65.01	35
LLM2Vec-Llama3-unsupervised	Llama-3-8B	Mean pool	Bidirectional	2 x 0.12M	56.23	139

LLM2Vec

Three steps to adapt any LLM to an encoder model:

- Enable bidirectional attention by modifying the attention mask.
- PEFT fine-tune base-LLM using masked next token prediction (MNTP).
- PEFT fine-tune of MNTP-tuned model using unsupervised contrastive learning (SimCSE).



The three steps of applying LLM2Vec, reproduced from Behnam Ghader et al., 2024.[2]

At release, LLM2Vec achieved SOTA unsupervised performance, on a small training dataset, but was only trained and evaluated on English.

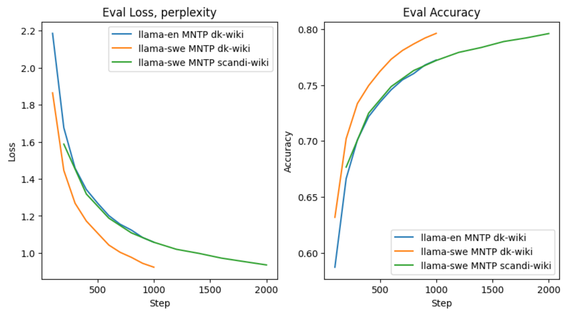
This project aims to explore the effectiveness of adopting the LLM2Vec approach to low-resource languages such as Danish, as well as explore the potential performance of a supervised-finetuning.

Key contributions

- Apply LLM2Vec on Danish to display the effect of fine-tuning on a low-resource language.
- Achieve SOTA unsupervised score on the Scandinavian embedding benchmark (SEB).
- Collect and share the largest combined dataset for supervised sentence-embedding finetuning in Danish (~100k samples).
- Achieve 7th best overall model on SEB by supervised finetuning on dataset using less than 1/10 train size of other top 10 models.

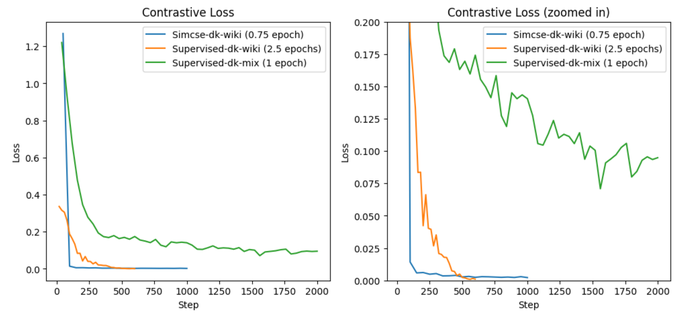
MNTP Training

Two base models: Llama3-en-instruct (Meta)[3] and a fine-tuned version from AI-sweden[4] . Fine-tuned on Danish Wikipedia[5] (*Similar to Behnam Ghader et al., 2024*) + Scandinavian Wiki[6].



Contrastive Training

Unsupervised: SimCSE on Danish Wikipedia[5] (sentences, *similar to Behnam Ghader et al., 2024*) .
Supervised: Multiple Negatives Ranking Loss (MNR) on synthetic Q&A dataset generated from DK Wiki[7] and on a new DK dataset curated from multiple sources including the aforementioned[8].

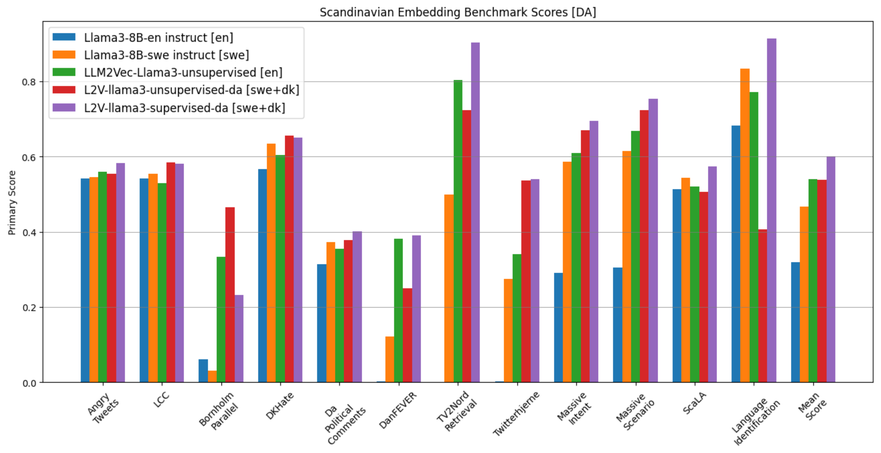


Results: Best-Performing (unsupervised) Embedding Model in Danish

Unsupervised model, **L2V-llama3-unsupervised-da**, scores 53.78 on SEB[da], beating current best model dfm-encoder-large-v1-SimCSE. Low performance on cross-lingual task *Language Identification* reduces overall mean score.

	Average Score	SEB rank	Embedding Size	Angry Tweets	Bornholm Parallel	DKHate	Da Political Comments	DanFEVER	LCC	Language Identification	Massive Intent	Massive Scenario	ScaLA	TV2Nord Retrieval	Twitterhjerne
Supervised models															
multilingual-e5-large-instruct (#1)	**66.4**	1	1024	**64.6**	**55**	67.1	**45.3**	39.5	**70.6**	82.5	**71.9**	**77.5**	50.2	93.7	77.2
e5-mistral-7b-instruct	61.7	4	4096	58.4	50.5	64.5	39.7	38.2	63.9	65.2	71.3	75.6	50.4	91.2	71.1
L2V-llama3-supervised-da	60.1		4096	58.21	23.28	65.05	40.03	39.06	58.00	**91.35**	69.40	75.37	**57.34**	90.20	53.91
text-embedding-3-small	59.7	7	1536	55.6	41	65.6	39.8	39.1	59.4	67.9	63.6	71.3	50.5	92	70.3
Self-supervised models															
L2V-llama3-unsupervised-da	**53.78**		4096	55.44	**46.59**	**65.53**	37.76	25.06	**58.47**	40.65	**66.97**	**72.35**	50.55	72.30	**53.68**
dfm-encoder-large-v1	47.70		1024	53.80	11.60	60.10	37.10	24.10	57.30	77.70	60.60	64.20	63.10	47.70	33.70
+ SimCSE (#1)	52.16		1024	54.42	15.93	63.19	**38.47**	**36.86**	58.07	75.98	65.83	71.61	**66.09**	**80.81**	16.99
XML Roberta	39.60		1024	51.70	4.30	60.20	31.90	10.60	48.70	**81.30**	47.30	49.50	60.30	6.10	20.40
XLM-roberta-base	38.10		768	52.40	4.40	56.80	33.70	8.70	52.30	79.40	41.10	43.90	57.30	5.30	18.80

Scores taken from the Scandinavian Embedding Benchmark and accompanying paper (Enevoldsen et al., 2024)[9]



Supervised model, **L2V-llama3-supervised-da**, scores 60.10 on SEB[da] ranking as overall 7th best model and outperforming OpenAI text-embedding-3-small.

Model has had very limited training data (100k) compared to the remainder of leaderboard that has >1M. Training curves indicate that the performance could be improved if more data was available.

References & data

1. Scores reported on the Hugging Face MTEB Leaderboard as of 8/12/2024
2. Parishad Behnam Ghader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024.Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.
3. Meta Llama 3 8B Instruct: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
4. AI Sweden, Llama 3 8B Instruct finetuned on Swedish: https://huggingface.co/AI-Sweden-Models/Llama-3-8B-instruct
5. Authors subset of Alexandra Institute cleaned Wikipedia Danish dataset: https://huggingface.co/datasets/jealk/wiki40b-da-clean
6. Authors subset of Alexandra Institute cleaned Wikipedia Scandinavian dataset: https://huggingface.co/datasets/jealk/scandi-wiki-combined
7. DDSC Synthetic queries generated from Danish Wikipedia articles (30k samples: https://huggingface.co/datasets/DDSC/da-wikipedia-queries-gemma-processed
8. Authors combined dataset of varied samples for MNR sentence training (100k samples: https://huggingface.co/datasets/jealk/supervised-da
9. Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. "The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding." arXiv:2406.02396 [cs.CL].