

Text Embedding Models

Frank Liu

A Quick Refresher



Embeddings models workhorses of AI apps



Embeddings models workhorses of AI apps

Hugging Face Q Search models, datasets, users							
sentence-transformers all-MiniLM-L6-v2							
Sentence Similarity Sentence Transformers 🍐 PyTorch 🏌 TensorFlow 😨 Rust 📑 s2orc							
 flax-sentence-embeddings/stackexchange_xml ms_marco gooaq yahoo_answers_topics code_search_net search_qa eli5 snli multi_nli wikihow natural_questions 							
<pre>trivia_qa = embedding-data/sentence-compression = embedding-data/flickr30k-captions</pre>							
embedding-data/SPECTER embedding-data/PAQ_pairs embedding-data/WikiAnswers (# English							
bert teature-extraction Inference Endpoints arxiv:1904.06472 arxiv:2102.07033 arxiv:2104.08727 arxiv:1704.05179 arxiv:1810.09305 <u>a</u> License: apache-2.0							
: % Deploy > Use in sentence-transformers							
Model card HE Files Community (2)							
∠ Edit model card							
i nis is a <u>sentence-transformers</u> model: it maps							





Back in the day...



 $W_{x,y} = tf_{x,y} \times log(\frac{N}{df_x})$





Back in the day...



 $W_{x,y} = tf_{x,y} \times log(\frac{N}{df_x})$

"Handcrafted features"





Modern day embeddings



Source: Arize Phoenix



Generating Text Embeddings



Recurrent neural networks



Source: CS230 notes



Recurrent neural networks



Source: CS230 notes



Transformer encoder



Source: <u>Illustrated Transformer</u>



Transformer encoder



Source: Illustrated Transformer

One embedding per token!



BERT





BERT



Cross-encoder (one inference per query/document pair)



Sentence BERT





Sentence BERT









- Some flavor of SBERT
 - Self-supervised pre-training (masking + NSP)
 - Contrastive (regression) or triplet loss
 - Changes to model arch or training, e.g. sparse attention, masking entities, or a different objective function



- Some flavor of SBERT
 - Self-supervised pre-training (masking + NSP)
 - Contrastive (regression) or triplet loss
 - Changes to model arch or training, e.g. sparse attention, masking entities, or a different objective function
- Highly data-dependent
 - Symmetric or asymmetric embeddings
 - Different models for different "domains" of text



- Some flavor of SBERT
 - Self-supervised pre-training (masking + NSP)
 - Contrastive (regression) or triplet loss
 - Changes to model arch or training, e.g. sparse attention, masking entities, or a different objective function
- Highly data-dependent
 - Symmetric or asymmetric embeddings
 - Different models for different "domains" of text
- Limited token length
 - Sweet spot seems to be around 100 200 tokens
 - Roughly equal to one paragraph



Which embedding models are best?

Rank 🔺	Model	Average 🔺	AskUbuntuDupQuestions	MindSmallReranking	SciDocsRR 🔺	StackOverflowDupQuestions
1	e5-mistral-7b-instruct	60.21	66.98	32.6	86.33	54.91
2	ember-v1	60.04	64.46	32.27	87.56	55.85
3	bge-large-en-v1.5	60.03	64.47	32.06	87.63	55.95
4	UAE-Large-V1	59.88	64.2	32.51	87.49	55.32
5	<u>sf_model_e5</u>	59.86	64.32	32.27	87.47	55.4
6	voyage-lite-01-instruct	59.74	65.77	31.69	87.03	54.49
7	all-mpnet-base-v2	59.36	65.85	30.97	88.65	51.98
8	<u>gte-large</u>	59.13	63.06	32.63	87.2	53.63
9	bge-base-en-v1.5-quant	58.94	62.39	31.89	87.05	54.45
10	<u>bge-base-en-v1-5-seqlen-384-bs-1</u>	58.86	62.13	31.2	87.49	54.61
11	bge-base-en-v1.5	58.86	62.13	31.2	87.49	54.61
12	stella-base-en-v2	58.78	62.72	31.91	86.66	53.81

Source: MTEB Leaderboard



Which embedding models are best?

- Determine your application requirements
 - What task am I interested in, e.g. classification, retrieval, etc?
 - Do we have enough data to fine-tune a model? \rightarrow make your own
 - How much latency are we okay with? \rightarrow pick smaller model
 - Are we okay with false positives? → specialized model



Which embedding models are best?

- Determine your application requirements
 - What task am I interested in, e.g. classification, retrieval, etc?
 - Do we have enough data to fine-tune a model? \rightarrow make your own
 - How much latency are we okay with? \rightarrow pick smaller model
 - Are we okay with false positives? → specialized model
- There is no one right answer

Demo Time

