Domain-Specific NLP

Introduction

Pretrained (Large) Language Models are trained on content crawled over the internet, books, reports and news papers and are, hence **are open-domain**.

A **textual domain** is the **distribution over language characterizing a given topic or genre** [1].

- You are more likely to see the word "integer" in computer science than in news papers.

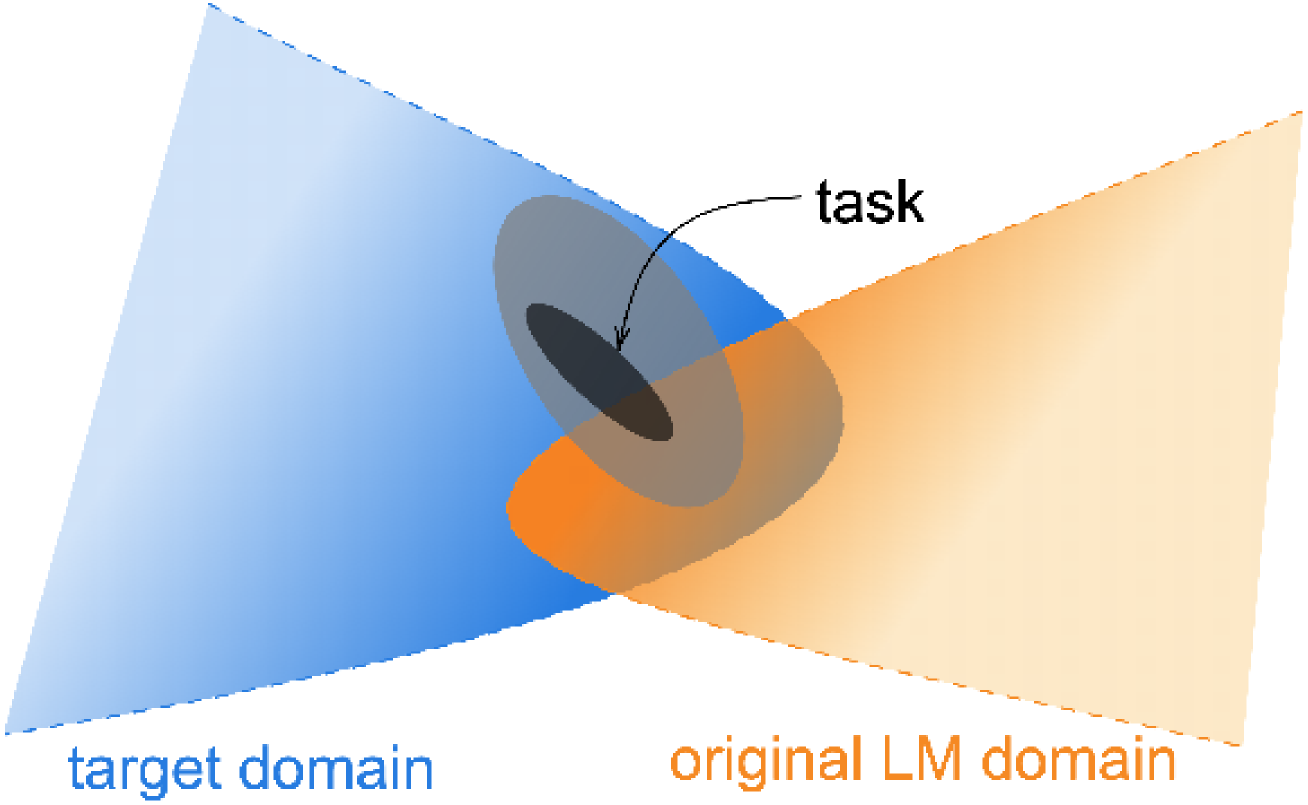
Contents

1. **Domain-Specific Models**
    a. *Don't stop pre-training*
    b. Specialized models (BioBERT, SciBERT, Galactica)
2. **Unsupervised Classification Models**
    a. Représentations out-of-the-box: limitations
    b. SimCSE, E5, GTE...
3. **Learning Long-Range Dependencies**
    a. Long-range attention models
    b. State-space models: S4

Domain-Specific Models

Don’t stop pre-training

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. [1]



Don't stop pre-training

Domain	Pretraining Corpus	# Tokens	Size	$\mathcal{L}_{\text{RoB.}}$	$\mathcal{L}_{\text{DAPT}}$
BIOMED	2.68M full-text papers from S2ORC (Lo et al., 2020)	7.55B	47GB	1.32	0.99
CS	2.22M full-text papers from S2ORC (Lo et al., 2020)	8.10B	48GB	1.63	1.34
NEWS	11.90M articles from REALNEWS (Zellers et al., 2019)	6.66B	39GB	1.08	1.16
REVIEWS	24.75M AMAZON reviews (He and McAuley, 2016)	2.11B	11GB	2.10	1.93
ROBERTA (baseline)	see Appendix §A.1	N/A	160GB	‡1.19	-

Table 1: List of the domain-specific unlabeled datasets. In columns 5 and 6, we report ROBERTA's masked LM loss on 50K randomly sampled held-out documents from each domain before ($\mathcal{L}_{ROB.}$) and after ($\mathcal{L}_{DAPT}$) $DAPT$ (lower implies a better fit on the sample). ‡ indicates that the masked LM loss is estimated on data sampled from sources similar to ROBERTA's pretraining corpus.

Don't stop pre-training

Figure 2: Vocabulary overlap (%) between domains. PT denotes a sample from sources similar to ROBERTA's pretraining corpus. Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords) in documents sampled from each domain.

Don't stop pre-training

Domain	Task	RoBERTa	Additional Pretraining Phases DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

Table 5: Results on different phases of adaptive pretraining compared to the baseline RoBERTa (col. 1). Our approaches are $DAPT$ (col. 2, §3), $TAPT$ (col. 3, §4), and a combination of both (col. 4).

Specialized models (BioBERT, SciBERT, Galactica)

"[..] the word distributions of general and biomedical corpora are quite different, which can often be a problem for biomedical text mining models." [2]

Specialized models (BioBERT, SciBERT, Galactica)

: 1. List of text corpora used for BioBERT

Corpus	# of words (B)	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical

Specialized models (BioBERT, SciBERT, Galactica)

"We showed that **pre-training BERT on biomedical corpora is crucial in applying it to the biomedical domain**. Requiring minimal task-specific architectural modification, **BioBERT outperforms previous models on biomedical text mining tasks** such as NER, RE and QA."

Specialized models (BioBERT, SciBERT, Galactica)

Task	Dataset	BioBERT	SciBERT
NER	BC5CDR	88.85	90.01
	JNLPBA	77.59	77.28
	NCBI-disease	89.36	88.57
REL	ChemProt	76.68	83.64

Table 2: Comparing SciBERT with the reported BioBERT results on biomedical datasets.

Specialized models (BioBERT, SciBERT, Galactica)

NB: SciBERT was trained on curated textual data ; not trained on code or script for example---at leat not trained directly and purposefully on this kind of data

Specialized models (BioBERT, SciBERT, Galactica)

"Unlike search engines, language models can potentially store, combine and reason about scientific knowledge." [4]

- Specialized models (BioBERT, SciBERT, Galactica) were trained on a rather small highly curated dataset.
- The data was standardized in markdown format.

Specialized models (BioBERT, SciBERT, Galactica)

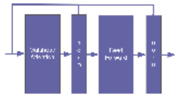
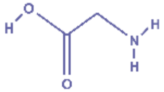
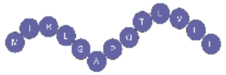
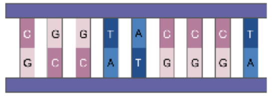
Modality	Entity	Sequence	
Text	Abell 370	`Abell 370 is a cluster...`	
$\LaTeX$	Schwarzschild radius	`r_{s} = \frac{2GM}{c^2}`	$r_s = \frac{2GM}{c^2}$
Code	Transformer	`class Transformer(nn.Module)`	
SMILES	Glycine	`C(C(=O)O)N`	
AA Sequence	Collagen α-1(II) chain	`MIRLGAPQTL..`	
DNA Sequence	Human genome	`CGGTACCCTC..`	

Table 1: Tokenizing Nature. Galactica trains on text sequences that represent scientific phenomena.

Table 1: Tokenizing Nature. Galactica trains on text sequences that represent scientific phenomena.

Specialized models (BioBERT, SciBERT, Galactica)

1. **Citations:** wrapped with special reference tokens [START_REF] and [END_REF].
2. **Step-by-Step Reasoning:** wrapped with a working memory token `<work>`, mimicking an internal working memory context.
3. **Mathematics:** for mathematical content, with or without LaTeX, ASCII operations are splitted into individual characters. Parentheses are treated like digits. The rest of the operations allow for unsplit repetitions. Operation characters are !"#$%&’*+,-./:;<=>?^_‘| and parentheses are ()[]{}.

4. **Numbers:** splitted into individual tokens. For example 737612.62 -> 7,3,7,6,1,2,.,6,2.
5. **SMILES formula:** wrapped with [START_SMILES] and [END_SMILES] and tokenized absed on characters. Similarly [START_I_SMILES] and [END_I_SMILES] is usedwhere isomeric SMILES is denoted.
6. **Amino acid sequences:** wrapped with [START_AMINO] and [END_AMINO] and apply character-based tokenization, treating each amino acid character as a single token. For example, MIRLGAPQTL -> M,I,R,L,G,A,P,Q,T,L.

1. **DNA sequences:** tokenized based on characters and wrapped inside [START_DNA] and [END_DNA]. For example, CGGTACCCTC -> C, G, G, T, A, C, C, C, T, C.

Specialized models (BioBERT, SciBERT, Galactica)

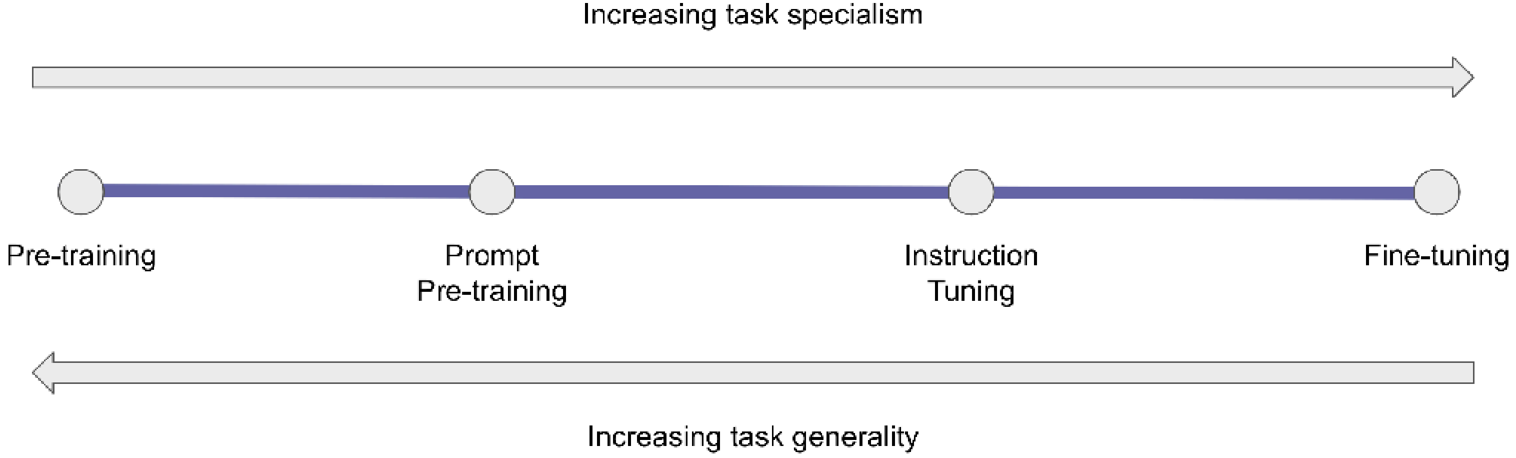


Figure 5: Prompt Pre-training. Pre-training weighs all tokens equally as part of the self-supervised loss. This leads to a weak relative signal for tasks of interest, meaning model scale has to be large to work. Instruction tuning boosts performance post hoc, and can generalize to unseen tasks of interest, but it risks performance in tasks that are distant from instruction set tasks. Prompt pre-training has a weaker task of interest bias than instruction tuning but less risk of degrading overall task generality.

Specialized models (BioBERT, SciBERT, Galactica)

- **GeLU Activation** - GeLU activations for all model sizes.
- **Context Window** - a 2048 length context window.
- **No Biases** - following PaLM, no biase in any of the dense kernels or layer norms.
- **Learned Positional Embeddings** - learned positional embeddings for the model.
- **Vocabulary** - vocabulary of 50k tokens using BPE. The vocabulary was generated from a randomly selected 2% subset of the training data.

Specialized models (BioBERT, SciBERT, Galactica)

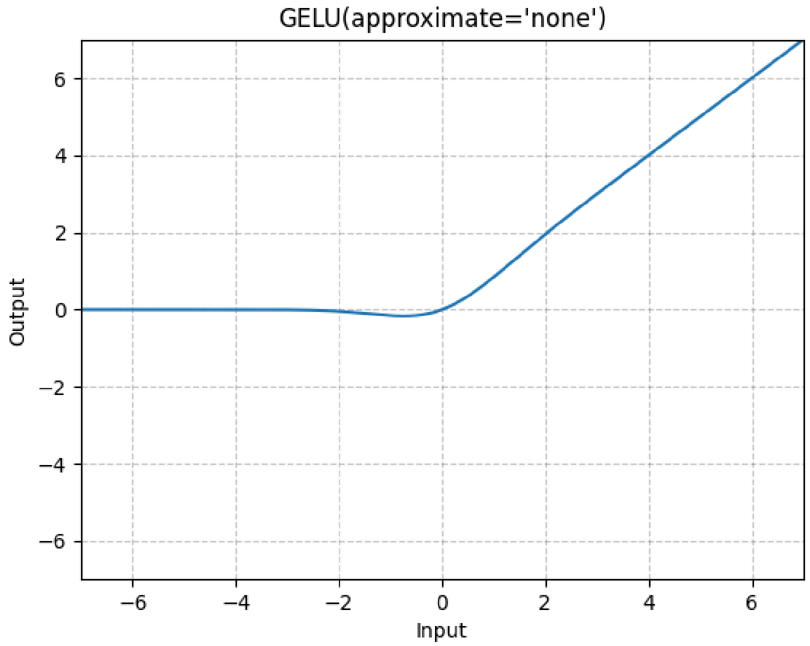
Gaussian Error Linear Units function (GeLu)

$$GELU(x) = x * \Phi(x)$$

Where $\Phi(x)$ is the Gaussian function.

$$GELU(x) \approx x * \frac{1}{2}(1 + Tanh(\frac{2}{\pi} * (x + 0.044715 * x^3)))$$

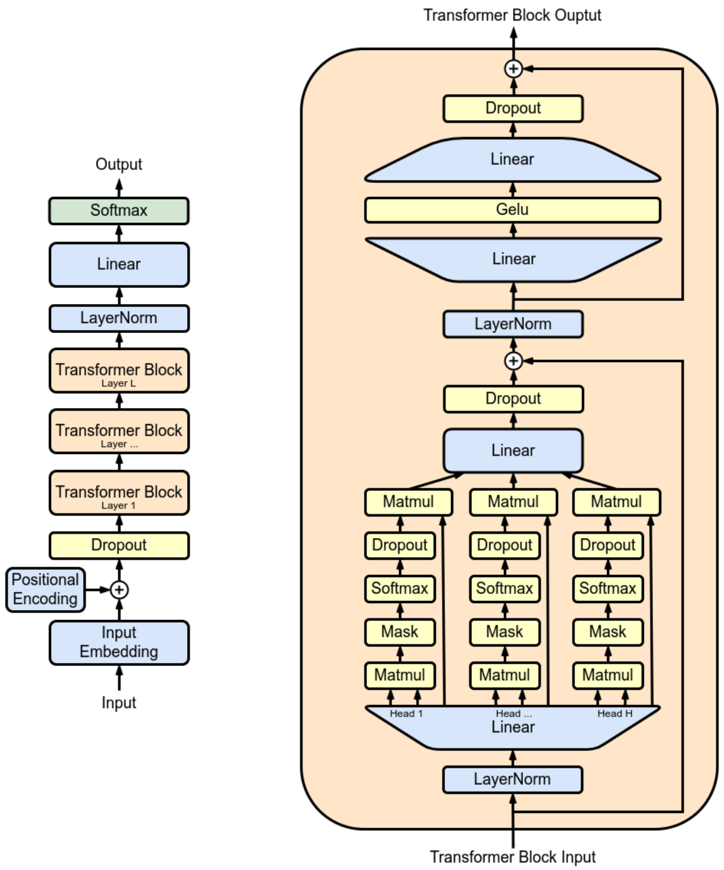
Specialized models (BioBERT, SciBERT, Galactica)



- Allows small negative values when $x < 0$.
- Avoids the dying ReLU problem.

Specialized models (BioBERT, SciBERT, Galactica)

Why no biases?



Unsupervised Classification Models

Représentations out-of-the-box: limitations

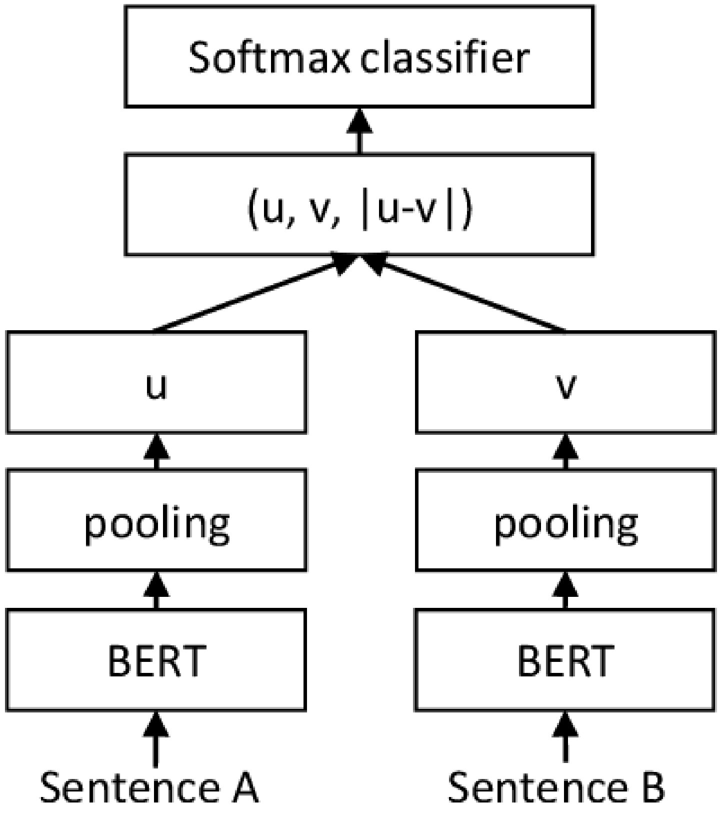
Embedding **pooling** is the process of **combining token embeddings** from an encoder model **into a single vector representing the entire input sequence**. Common methods include averaging (**mean pooling**), taking the maximum (**max pooling**), or using a **special token** like `[CLS]` or `<s>`.

Représentations out-of-the-box: limitations

Représentations out-of-the-box: limitations

Représentations out-of-the-box: limitations

Représentations out-of-the-box: limitations



NB: when your output pooled vectors are scaled, the cosine similarity is equal to the dot product.

Représentations out-of-the-box: limitations

Late interaction:

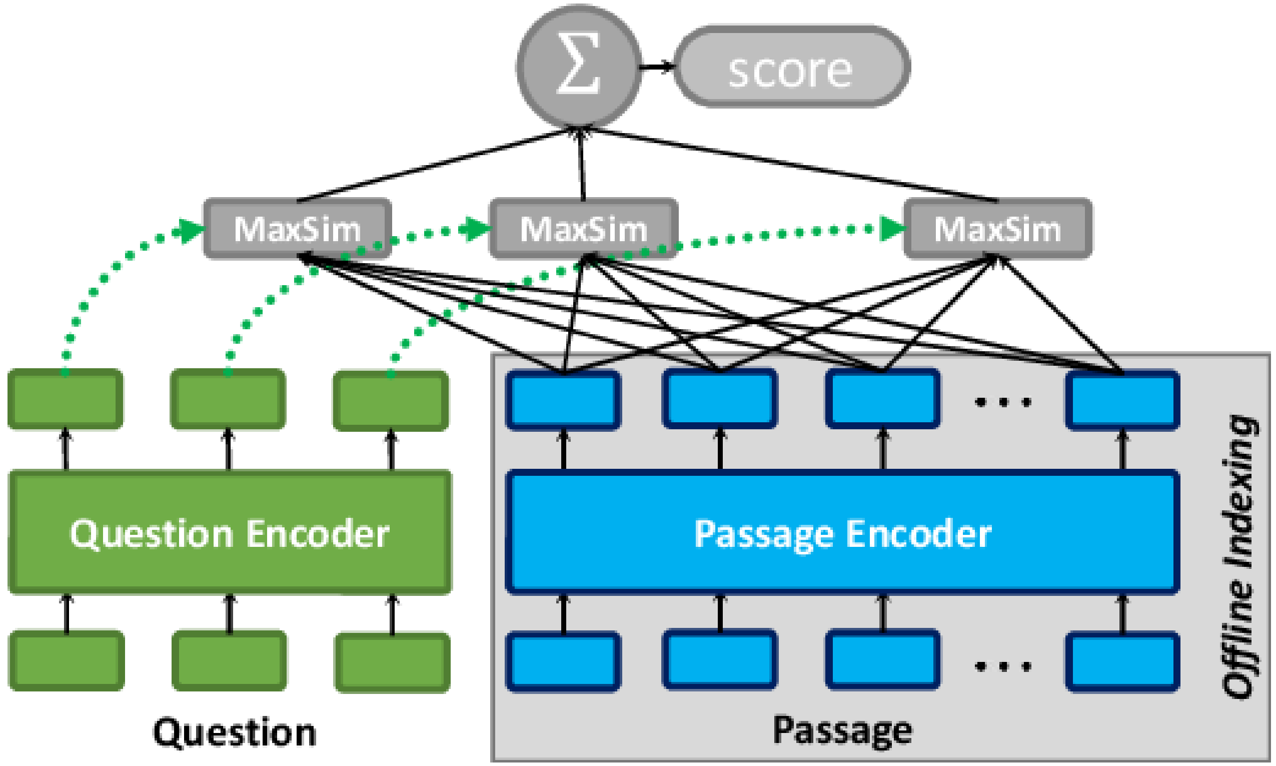


Figure 1: The late interaction architecture, given

Représentations out-of-the-box: limitations

The data is being compressed mutliple time -> challeging document can be hard to embed.

Can we do better?

SimCSE, E5, GTE...

Contrastive learning uses **similar data point** and **opposite ones** in order for the model build **close representations for the first ones** and and **more separated ones for the latter**. [7]

- Unsupervised SimCSE: standard dropout as data augmentation
- Supervised SimCSE: use pairs in NLI datasets

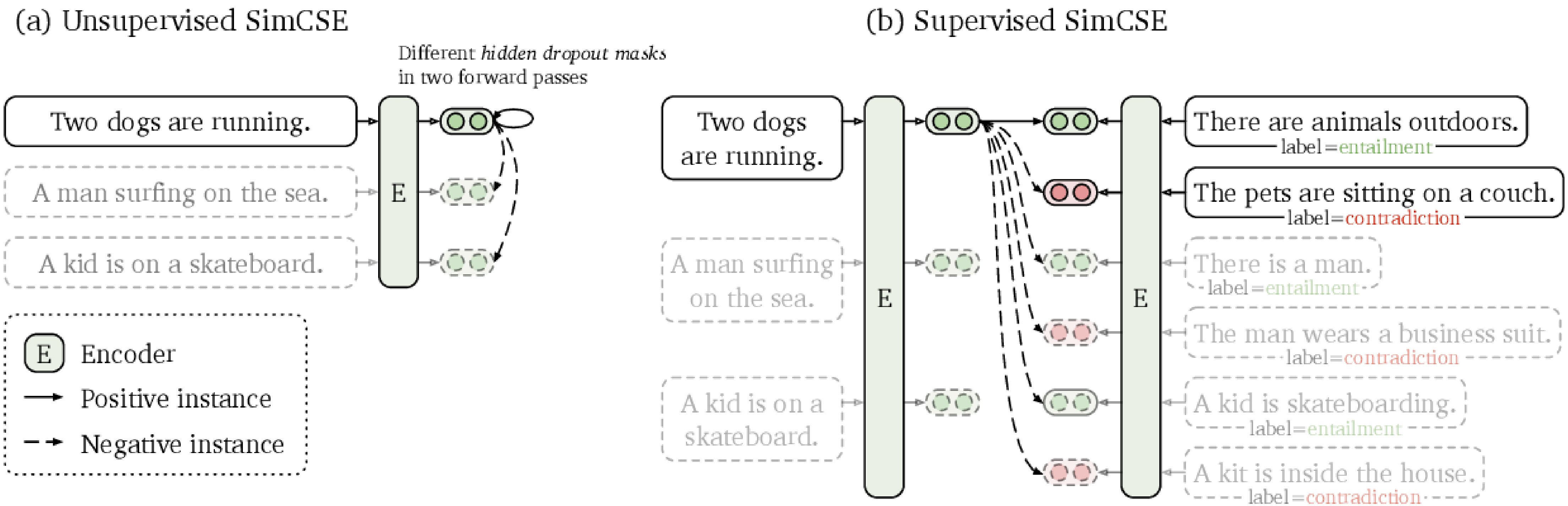
SimCSE, E5, GTE...

$$\mathcal{L}_{uns} = -log \frac{exp(\frac{sim(\mathbf{h}_i, \mathbf{h}_i^+)}{\tau})}{\sum_{j=1}^{N} exp(\frac{sim(\mathbf{h}_i, \mathbf{h}_j^+)}{\tau})}$$

$$\mathcal{L}_{sup} = -log \frac{exp(\frac{sim(\mathbf{h}_i, \mathbf{h}_i^+)}{\tau})}{\sum_{j=1}^{N} exp(\frac{sim(\mathbf{h}_i, \mathbf{h}_j^+)}{\tau}) + exp(\frac{sim(\mathbf{h}_i, \mathbf{h}_j^-)}{\tau})}$$

[8]

SimCSE, E5, GTE...



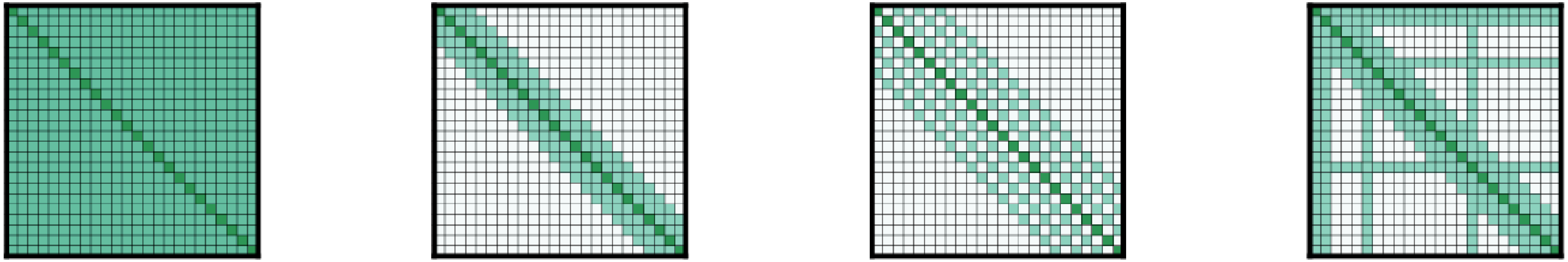
SimCSE, E5, GTE...

Contrastive learning mitigates anisotropy in language models by encouraging **embeddings** to be **more uniformly distributed** in the representation space. It pulls similar embeddings closer and pushes dissimilar ones apart, preventing over-clustering and ensuring better geometric properties for downstream tasks.

Learning Long-Range Dependencies

Long-range attention models

Sliding window attention: Longformer [11]



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Long-range attention models

Sliding window attention: Mistral 7B [12]

2 Architectural details

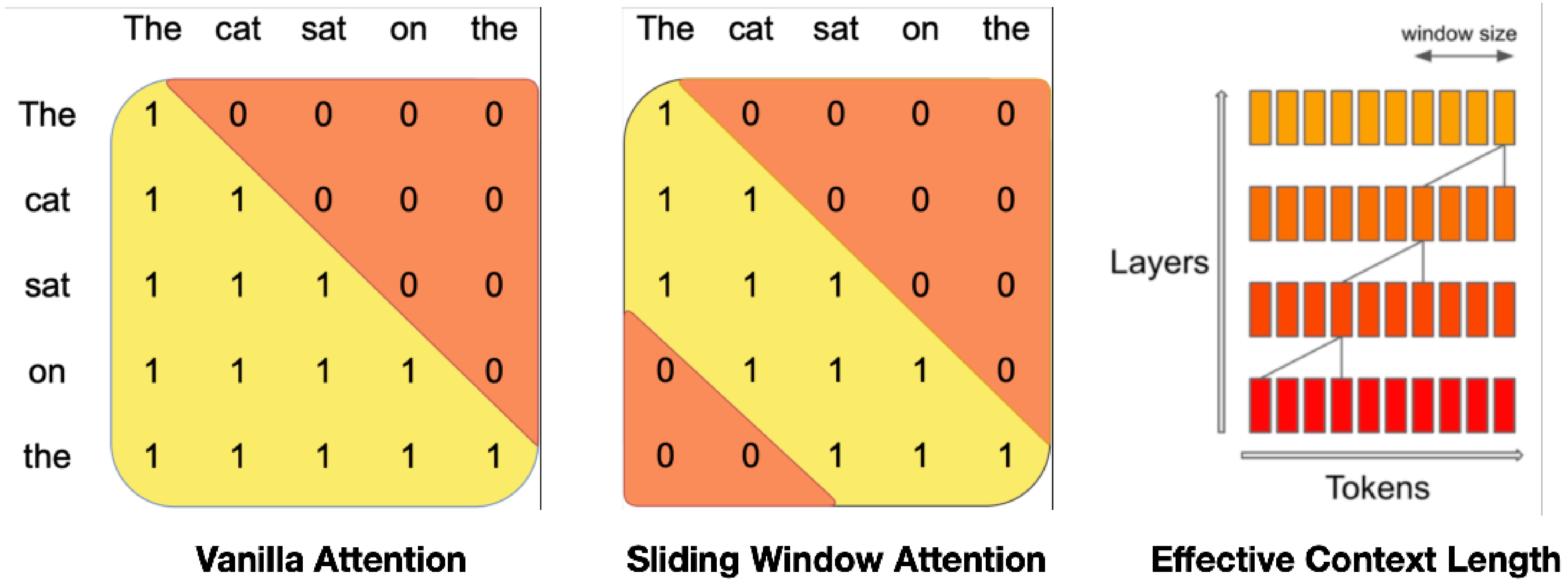
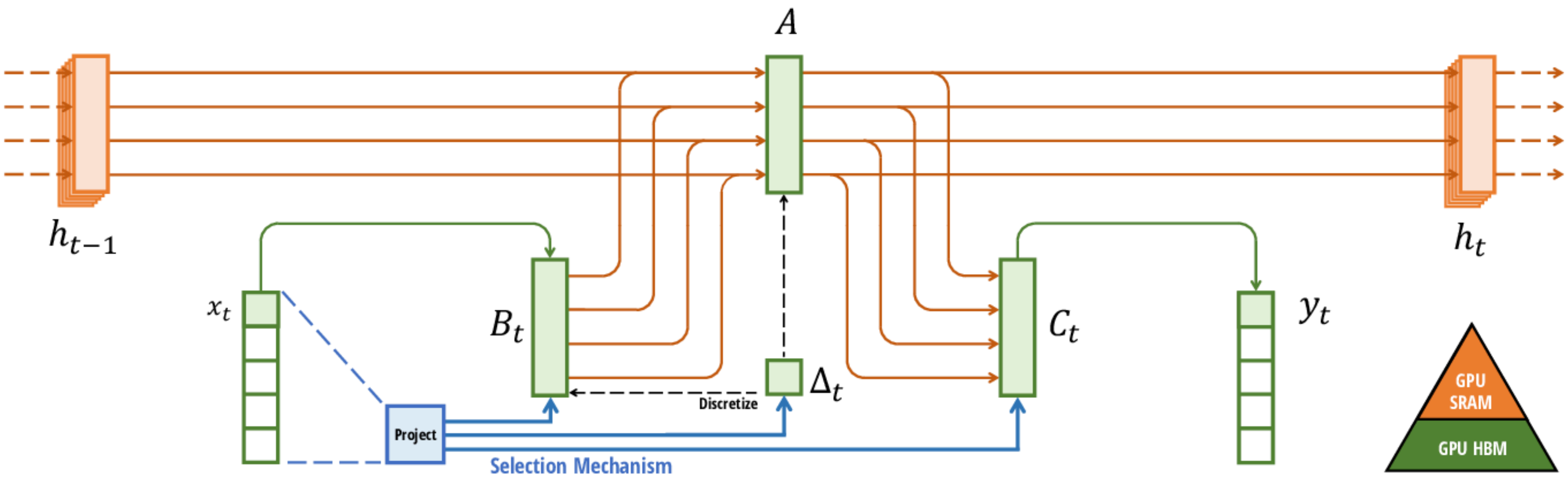


Figure 1: Sliding Window Attention. The number of operations in vanilla attention is quadratic in the sequence

State-space models: Mamba



Questions?

References

[1] Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 8342–60. Online: Association for Computational Linguistics, 2020.

[2] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." Bioinformatics 36, no. 4 (February 15, 2020): 1234–40.

[3] Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 3615–20. Hong Kong, China: Association for Computational Linguistics, 2019.

[4] Taylor, Ross, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. "Galactica: A Large Language Model for Science." arXiv, November 16, 2022.

[5] Nurmambetova, Elvira, et al. "Developing an Inpatient Electronic Medical Record Phenotype for Hospital-Acquired Pressure Injuries: Case Study Using Natural Language Processing Models." JMIR AI 2.1 (2023): e41264.

[6] Reimers, Nils, and Iryna Gurevych. "[Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.]()" In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 3982–92. Hong Kong, China: Association for Computational Linguistics, 2019.

[7] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 6894–6910. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021.

[8] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation Learning with Contrastive Predictive Coding." arXiv, January 22, 2019.

[9] Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. "Text Embeddings by Weakly-Supervised Contrastive Pre-Training." arXiv, December 7, 2022.

[10] Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. "Towards General Text Embeddings with Multi-Stage Contrastive Learning." arXiv, August 6, 2023.

[11] Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer." arXiv, December 2, 2020.

[12] Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. "Mistral 7B." arXiv, October 10, 2023.

[13] Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2021). Colbertv2: Effective and efficient retrieval via lightweight late interaction. arXiv preprint arXiv:2112.01488.

[14] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.