IDO / 2025
CONF

Международная научно-практическая конференция
ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА (NLP) ДЛЯ ЛИНГВИСТИКИ, IT И БИЗНЕСА

Malashenko Boris
ITMO University
quelquemath@gmail.com

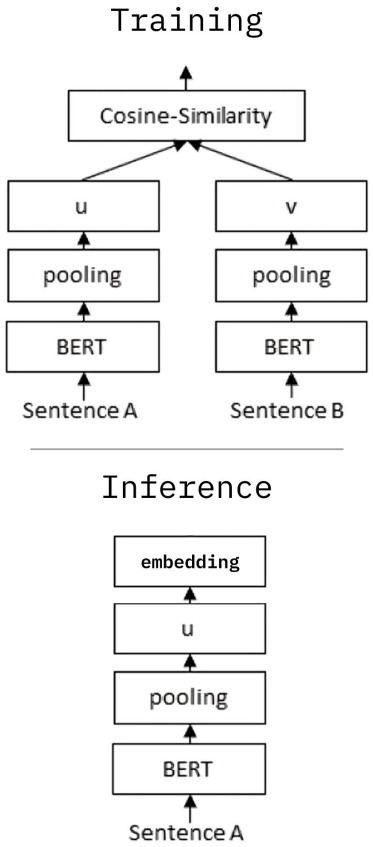
Text-level distillation: How to Build Small but Powerful Sentence Encoders

Sentence encoders: what are they?

Sentence encoders transform text into fixed-size vector representations that capture semantic meaning. These models are typically trained using contrastive learning or paired objectives on large text corpora.

Use cases:

- **Business:** Document clustering, customer support automation, and sentiment analysis.
- **Linguistics:** Analyzing semantic similarity between distinct languages, studying language patterns.

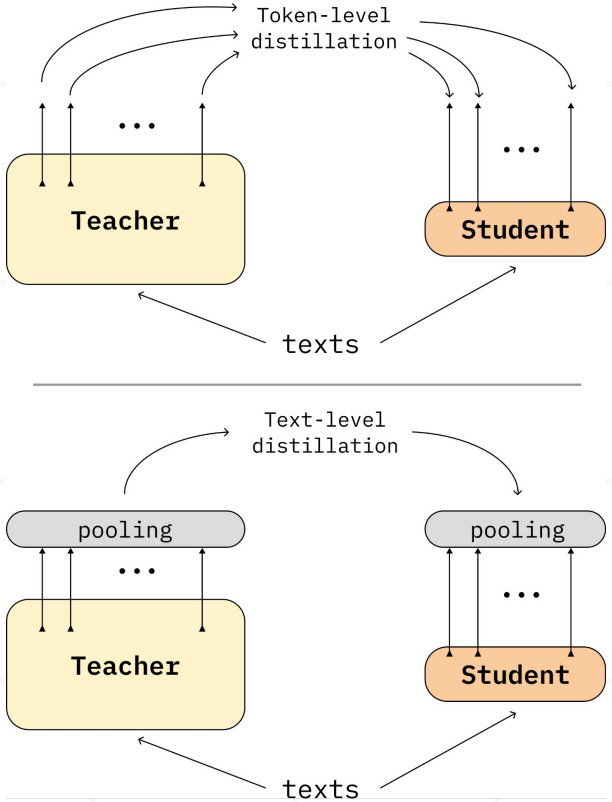


Text-level distillation

Distillation is a model compression technique where a smaller **student model** learns to replicate the behavior of a larger **teacher model** by minimizing the difference between their outputs.

For encoders like BERT, the student model typically shares the same tokenizer as the teacher, allowing it to learn by mimicking the teacher's token-level logits.

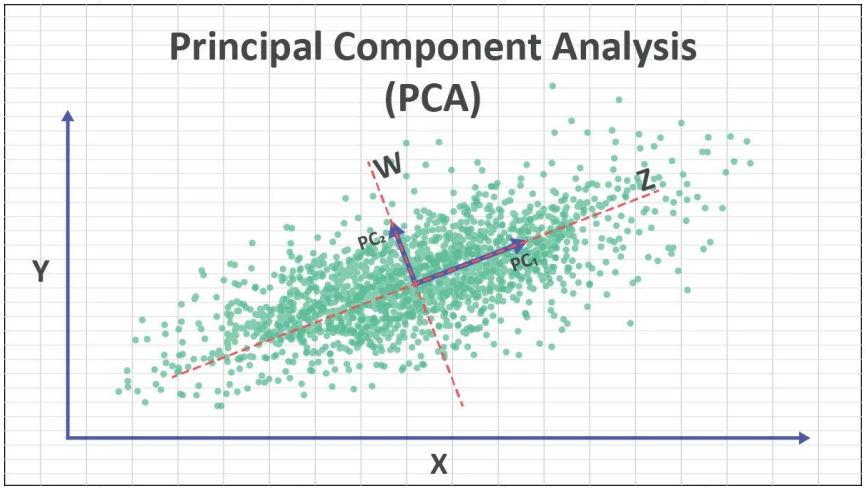
However, if the models use different tokenizers or the teacher's weights are unavailable (e.g., closed-source models), token-level distillation becomes impractical. Additionally, this approach does not directly optimize the quality of the final sentence embedding, which is crucial for sentence encoders.



Reducing dimensionality

The teacher and student models may have different hidden representation sizes, making direct alignment challenging.

To address this, we apply dimensionality reduction techniques such as Gaussian projection and PCA, which transform the teacher's embeddings into a lower-dimensional space compatible with the student.



Losses

While MSE loss is a viable option for aligning teacher and student representations, it can be modified in two key ways:

1) focus on cosine similarity, as it will be our target distance metric afterwards;

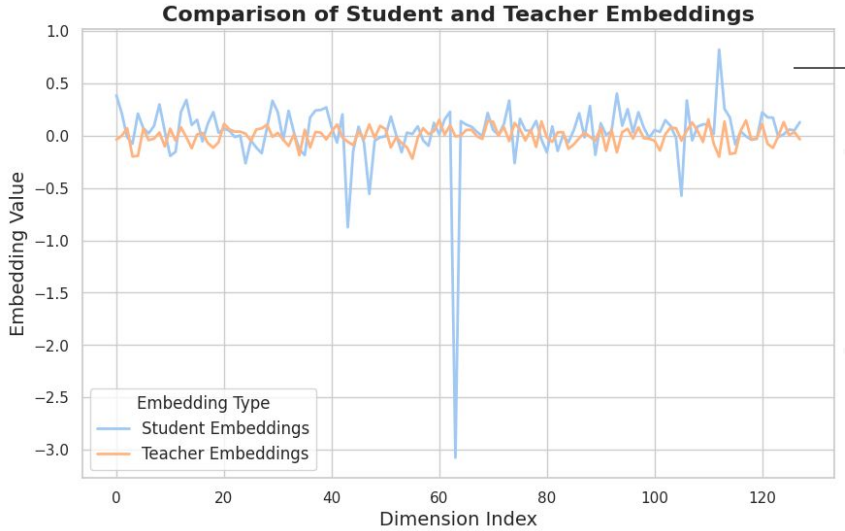
2) soft-targeting, to lower pre-train knowledge degradation.

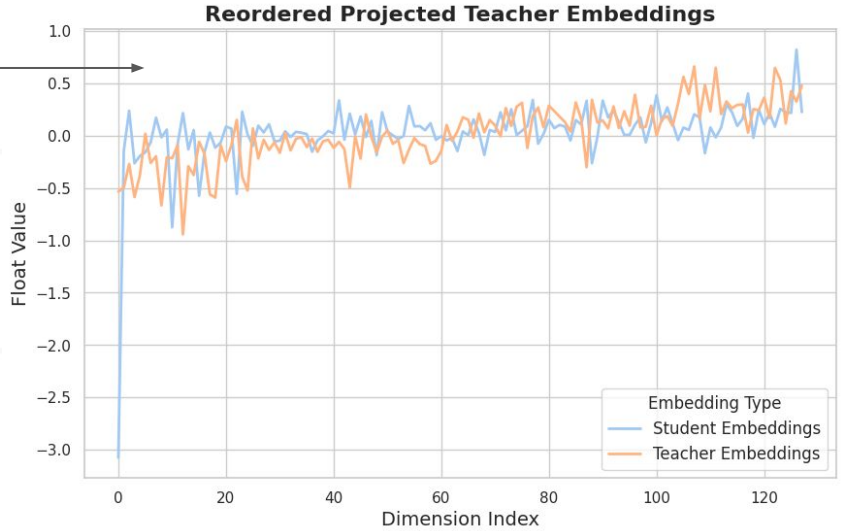
$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(t_i - s_i)^2$$

$$\mathcal{L}_{\text{SoftCos}} = \frac{1}{N}\sum_{i=1}^{N}\left|\left|\left|\left|\frac{\hat{t_i}+\hat{s_i}}{2}\right|\right| - \hat{s_i}\right|\right|^2$$

Latent space alignment

The teacher's initial latent space may differ significantly from the student's. Since it is fixed, we can apply linear transformations to align them.

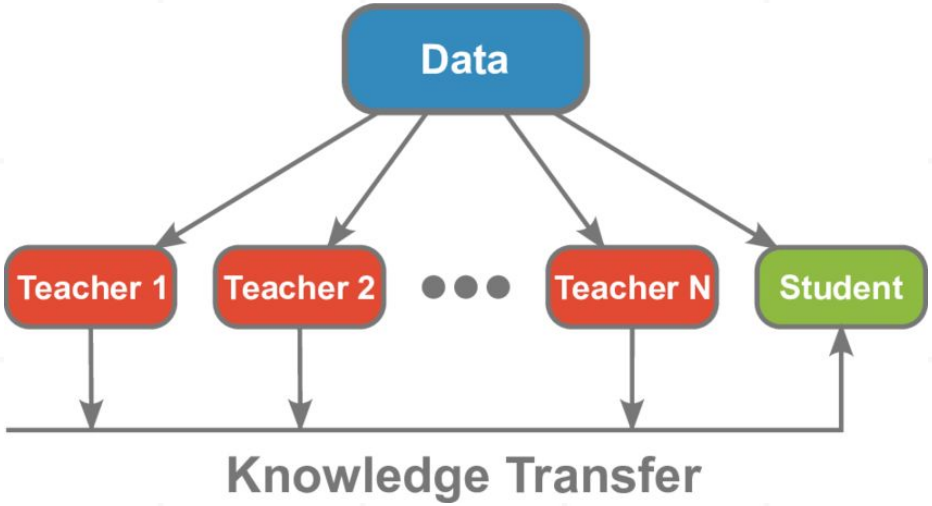




Multi-teacher distillation

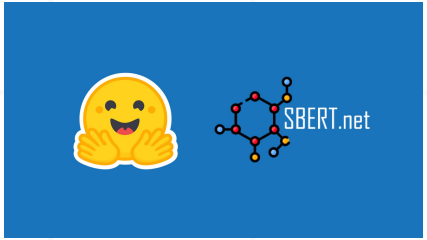
In multi-teacher distillation, knowledge is transferred from multiple teachers to a single student. This allows the student to learn diverse representations from different models.

One approach is **split distillation,** where different parts of the student model are trained using different teachers.



Training setup

Architecture: I use the tiny BERT model (google/bert_uncased_L-2_H-128_A-2), the smallest in the series with only 4.5 million parameters.



Tokenizer & Pre-training: As the model was originally trained on English, I trained a new tokenizer on a subset of the deepvk/cultura_ru_edu corpus. Subsequently, I conducted a pre-training on the fineweb-v2 dataset. After distillation, a brief fine-tuning phase was carried out.

Framework & Teacher: Experiments are performed using the Sentence Transformers framework, with the bge-m3 model serving as the teacher.

Training Settings: The training configuration includes 200 warmup steps, a batch size of 512, 10,000 total training steps, and a weight decay of 0.01.

Results: reducing dimensionality

- **STS** (Semantic Text Similarity): Tasks that assess the semantic similarity between text pairs.
- **IR** (Information Retrieval): Techniques focused on retrieving relevant information from large datasets.

Our experiments indicate that both PCA and Gaussian Random Projection yield top performance. PCA excels in STS tasks by preserving principal components, while GRP modifies distances yet retains essential information, benefiting IR tasks.

	STS	IR
Teacher	0.863	0.726
Student	0.650	0.140
Dimension reduction method	**STS**	**IR**
RP: Gaussian	0.840	**0.562**
PCA	**0.849**	0.560
Ridge	0.836	0.388
CCA	0.847	0.524

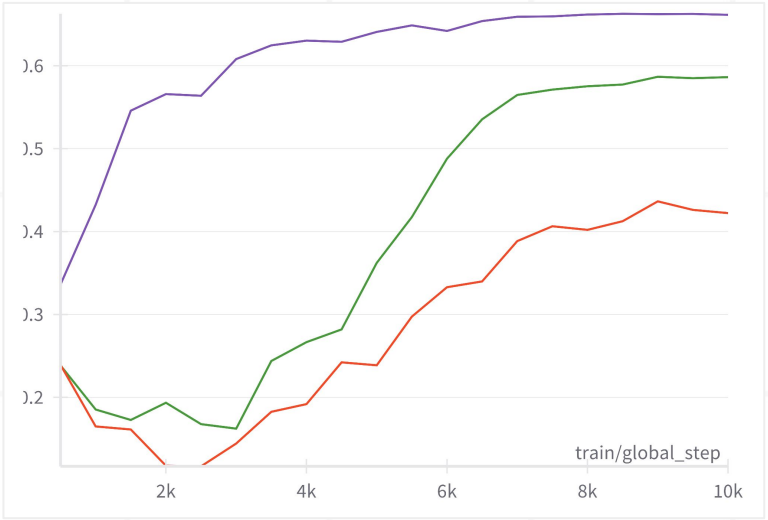
Results: training

Red – MSE

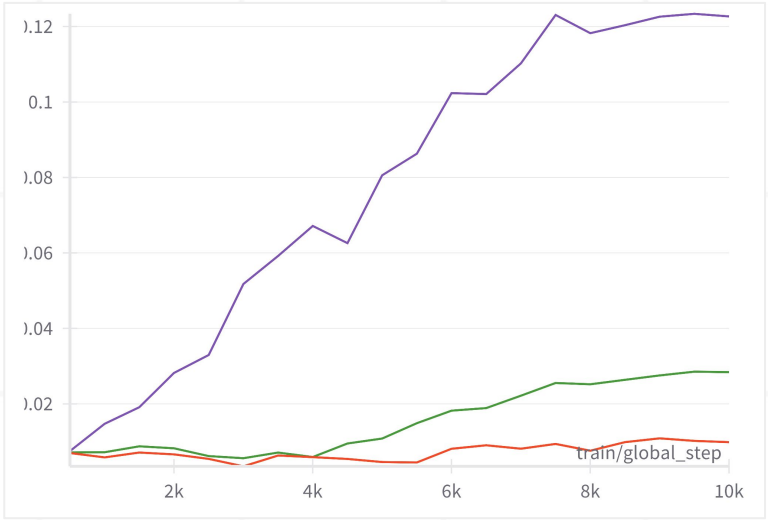
Green – SoftCosMSE

Violet – SoftCosMSE + space alignment

STS



IR



MTEB Results

All models were evaluated using the Russian subset of the MTEB benchmark, which comprises 21 tasks; However, I ran only 18, excluding the most resource-intensive ones: Miracl, RiaNews.

Model	Size	MTEB score
rubert-tiny	12M	0.396
rubert-tiny2	29M	0.469
google_tiny+cos+reordering	4.5M	0.428
google_tiny+cos+reordering+sft	4.5M	**0.478**

Malashenko Boris
quelquemath@gmail.com
tg: btmalov

Thank you for your attention!