EMBED YOUR DATA FOR FREE OPENSOURCE EMBEDDING MODELS

BY "REIMERS, NILS AND GUREVYCH, IRYNA",

arXiv

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) has set a new state-of-the-art performance on sentence-pair regression tasks like semantic textual similarity (STS). However, it requires...

arXiv.org

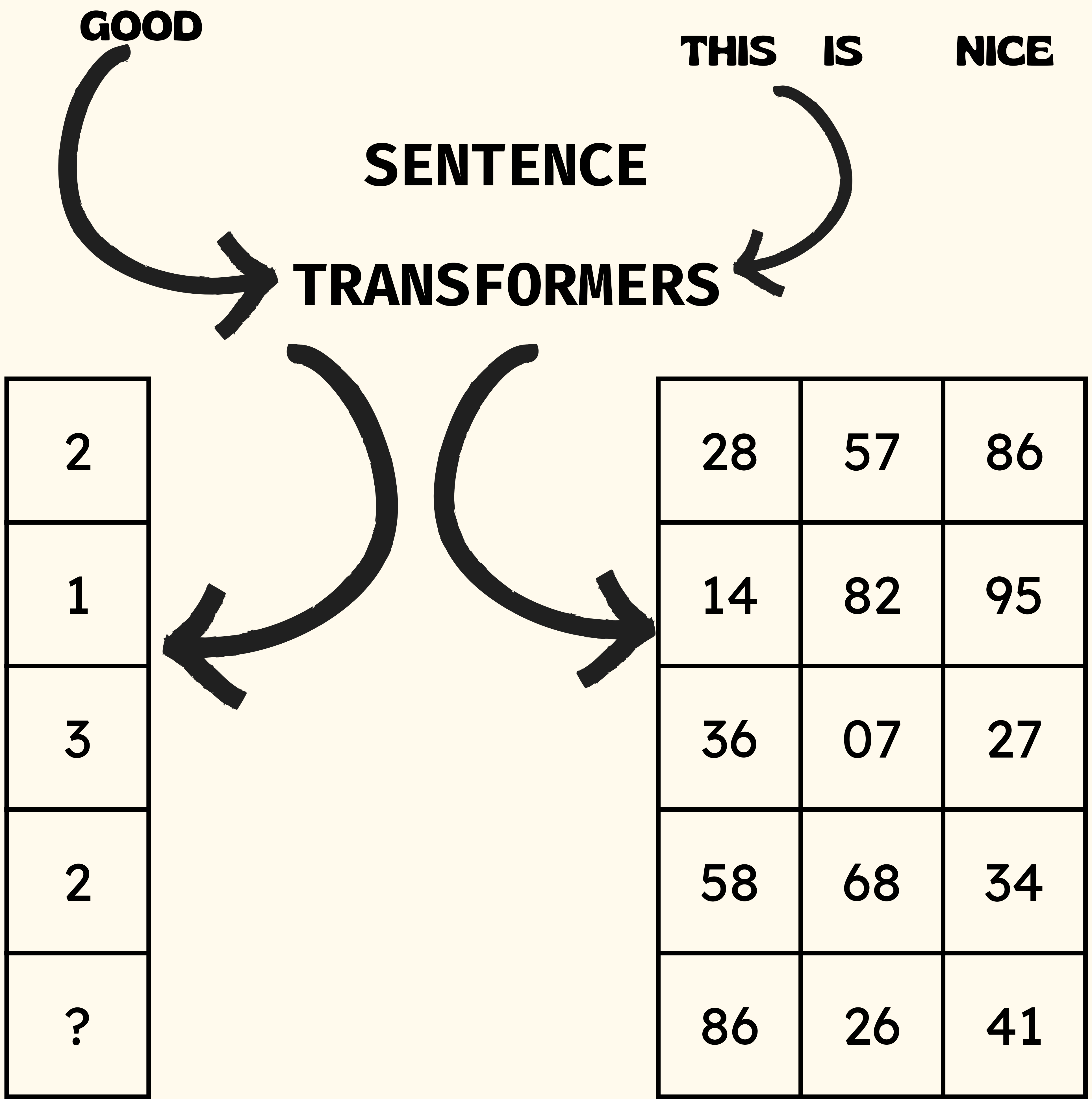
DEEP DIVE INTO CONCEPTS & CODE

SBERT.net

HTTPS://GITHUB.COM/INSIGHTBUILDER

CHALLENGE SOLVED: DATA PROCESSING FOR MODELING

- SENTENCE TRANSFORMERS ARE MODELS THAT CONVERT TEXT TO VECTORS OF NUMBERS
- USAGE:
  - SEMANTIC SIMILARITY
  - SEMANTIC SEARCH
  - RETRIEVE & RE-RANK
  - CLUSTERING
  - PARAPHRASE MINING
  - TRANSLATED SENTENCE MINING
  - CROSS ENCODERS
  - IMAGE SEARCH



HTTPS://GITHUB.COM/INSIGHTBUILDER

USAGE CONCEPTS: HOW EMBEDDINGS ARE USED?

EMBEDDINGS

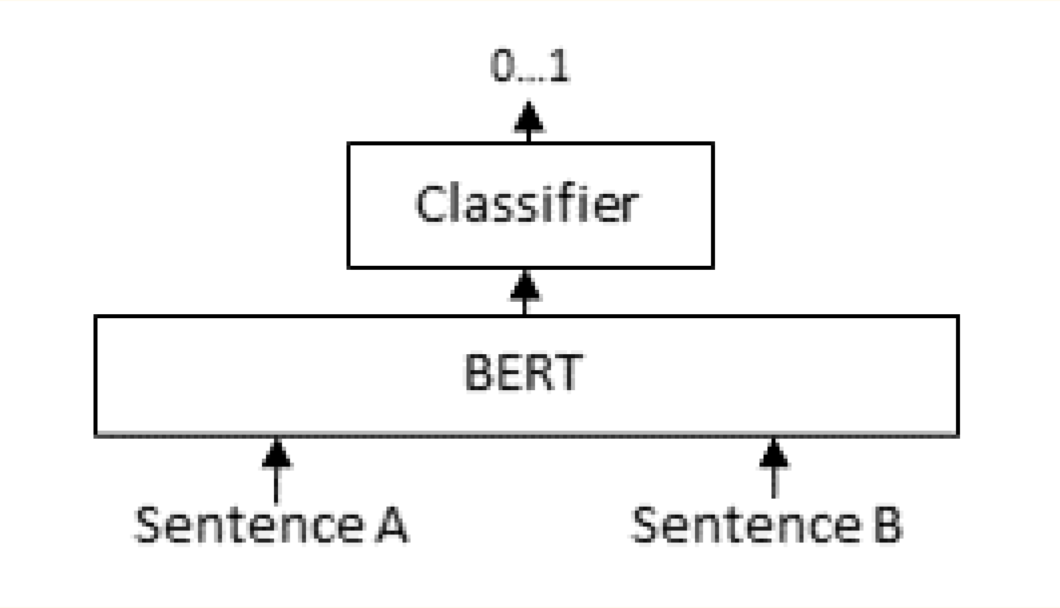
28	57	86
14	82	95
36	07	27
58	68	34
86	26	41

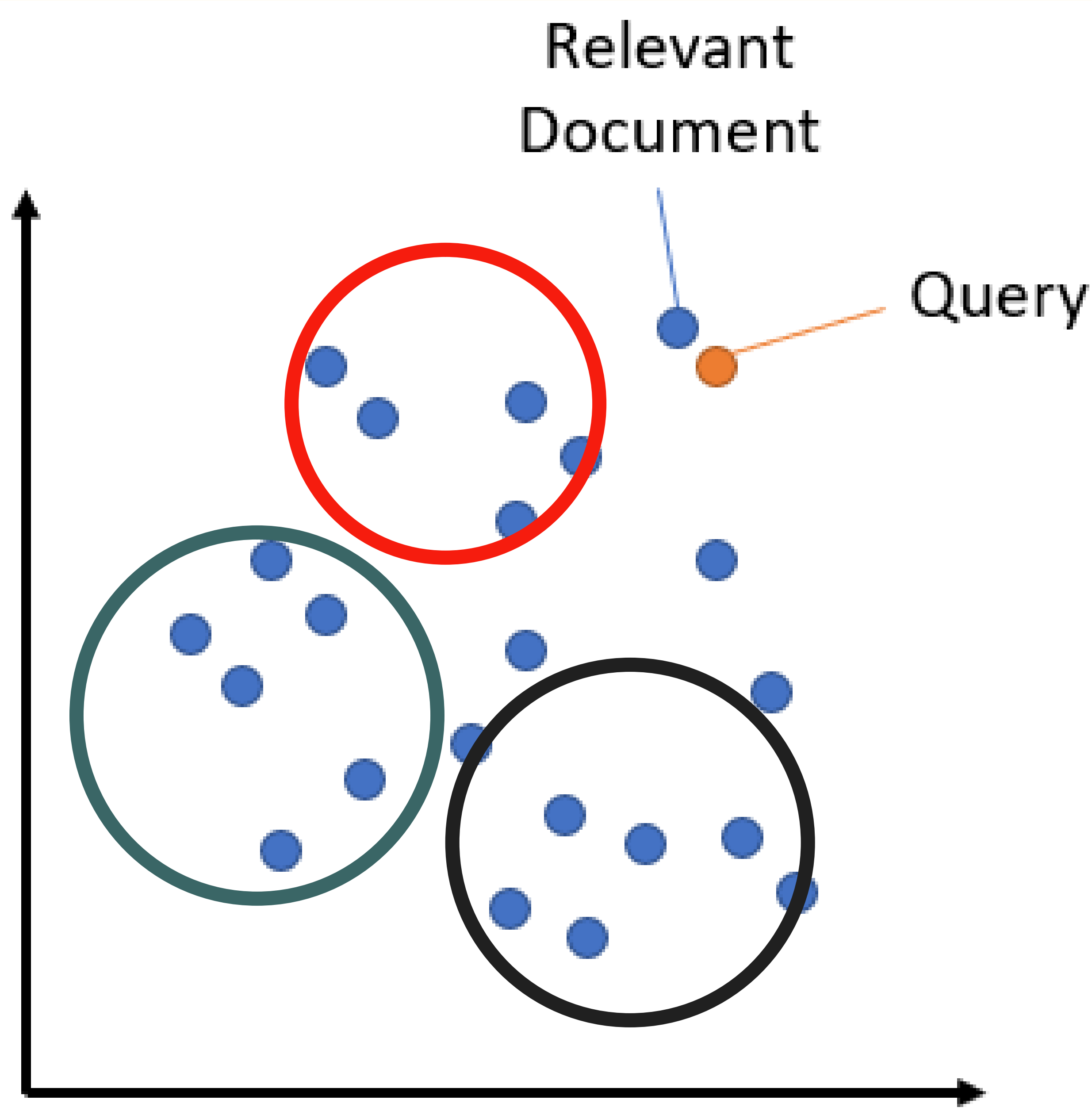
QUERY EMBED

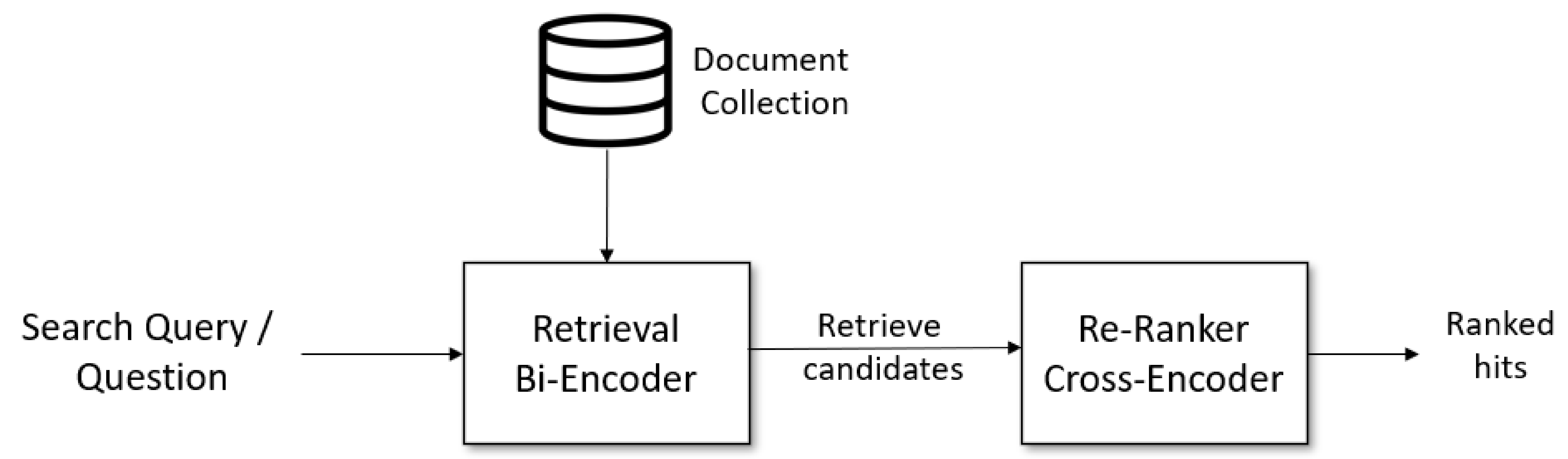
2
1
3
2
?

COSINE SIMILARITY

- **NEARER TWO SENTENCES SIMILAR THEY ARE**
- **PARAPHRASE MINING IS USED WHEN MULTIPLE SENTENCE TO BE CHECKED**
- **SEMANTIC SEARCH IS DONE BY POSITIONING THE QUERY IN VECTOR SPACE AND CHECKING THE NEAREST NEIGHBOURS**
- **SYMMETRIC VS ASSYMMETRIC SEARCH**
- **RETRIEVE AND RE-RANK WITH CROSS ENCODER SEARCH**
- **FAST / AGGLOMERATIVE CLUSTERING / TOPIC MODELING**







CONCEPTS DEEP DIVE: EXPLAINING APPLICATION

- SIMILAR SENTENCE : DISTANCE METRICS IS USED FOR FINDING SIMILAR SENTENCES
- PARAPHRASE MINING: QUERY IS COMPARED WITH CHUNKED CORPUS RATHER THAN SINGLE SENTENCES
- SEMANTIC SEARCH : QUERY IS EMBEDDED AND THEN PLACED IN VECTOR SPACE AS THAT OF THE CORPUS. ALL NEAREST SENTENCES ARE CONSIDERED SIMILAR
- RE-RANKING: ON TOP OF RETRIEVING THE SENTENCES, EACH OF THE SENTENCES IS CHECKED USING A CROSS ENCODER WHICH CLASSIFIES SENTENCE AT RAPID RATE
- BI-ENCODER PRODUCES SENTENCE EMBEDDINGS AND THEN COMPARES, WHILE THE CROSS ENCODER DOES NOT PRODUCE EMBEDDINGS, AND DIRECTLY CLASSIFIES THE SENTENCES

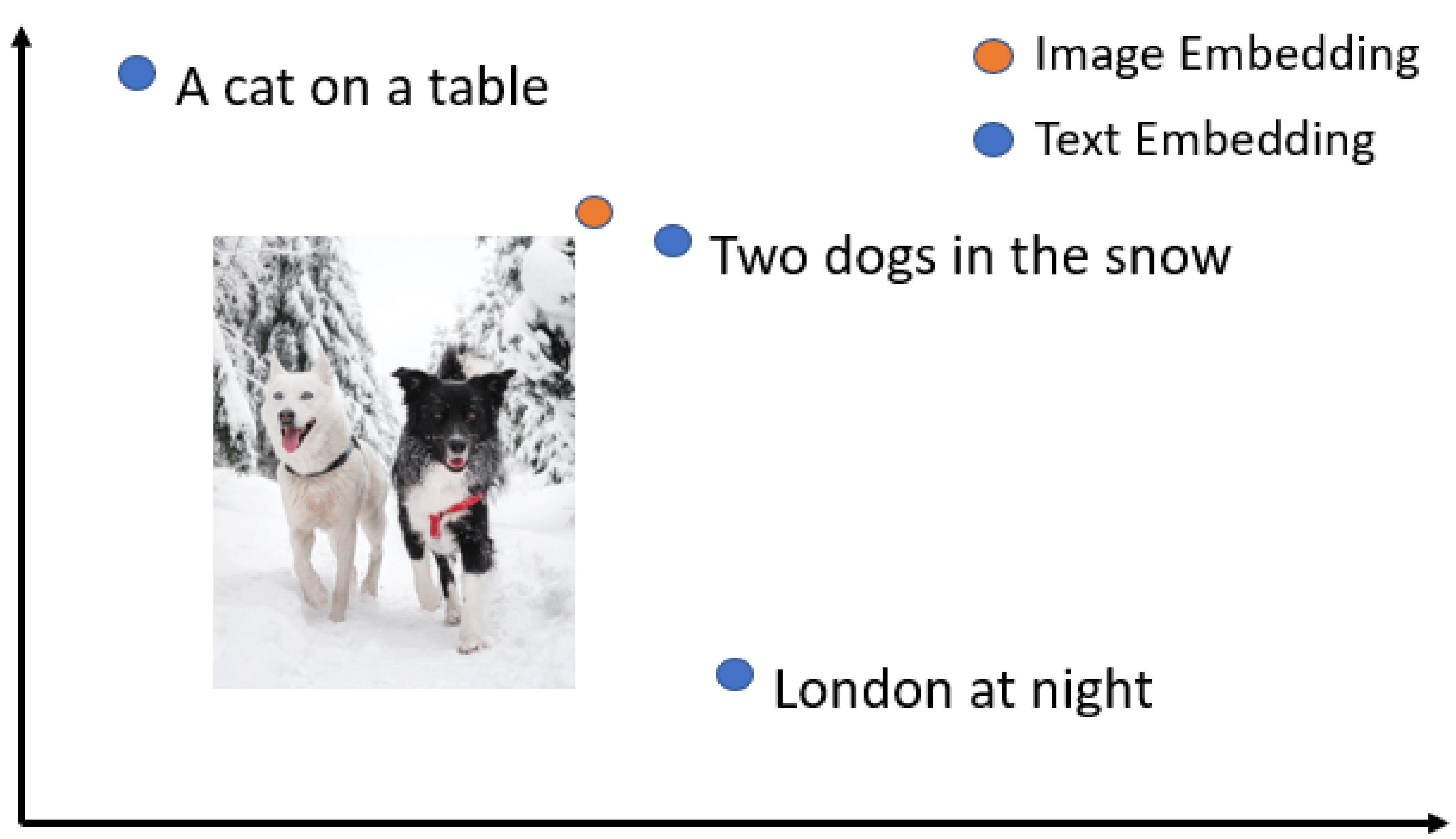


IMAGE TO TEXT SIMILARITY

CLIP MODEL FOR:

- TEXT-TO-IMAGE / IMAGE-TO-TEXT / IMAGE-TO-IMAGE / TEXT-TO-TEXT SEARCH
- YOU CAN FINE-TUNE IT ON YOUR OWN IMAGE&TEXT DATA

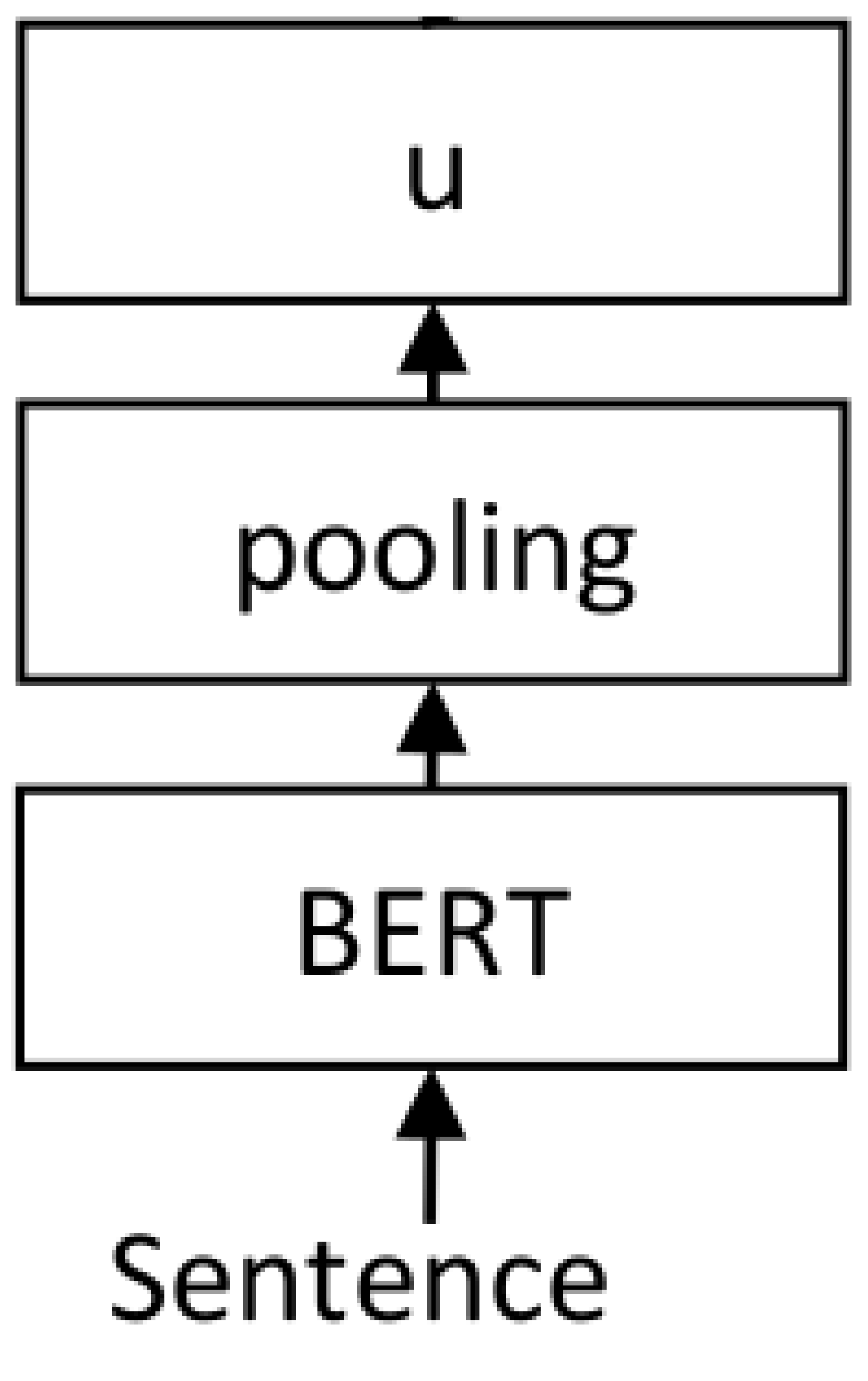
CUSTOM EMBEDDING : TRAINING ON OWN TEXT / IMAGE

WHY FINE-TUNING?

EMBEDDING FOR SPECIFIC TASK IMPROVES THE PERFORMANCE. THE TRAINING STRATEGY DEPENDS ON THE TASK AND THE DATA FORMAT.

PROCESS OF TUNING THE NEURAL NET

- DECIDE THE SENTENCE LENGTH
- CREATING MODEL ARCHITECTURE FROM SCRATCH USING MODULES
- DATA FOR TRAINING IS FORMULATED BASED ON THE TASK TO BE PERFORMED BY THE MODEL
- INPUT EXAMPLES WILL HAVE THE TEXT DATA ALONG WITH THE LABEL.
- LOSS FUNCTION IS ALSO DECIDED BASED ON THE TASK AND THE AVAILABLE DATASET
- EVALUATION IS PROCESS OF CHECKING MODEL PERFORMANCE IN REALITY
- MODEL.FIT(TRAIN_OBJECTIVES, EVALUATOR, EVALUTION_STEPS, EPOCHS, WARMUP_STEPS, OUTPUT_PATH)



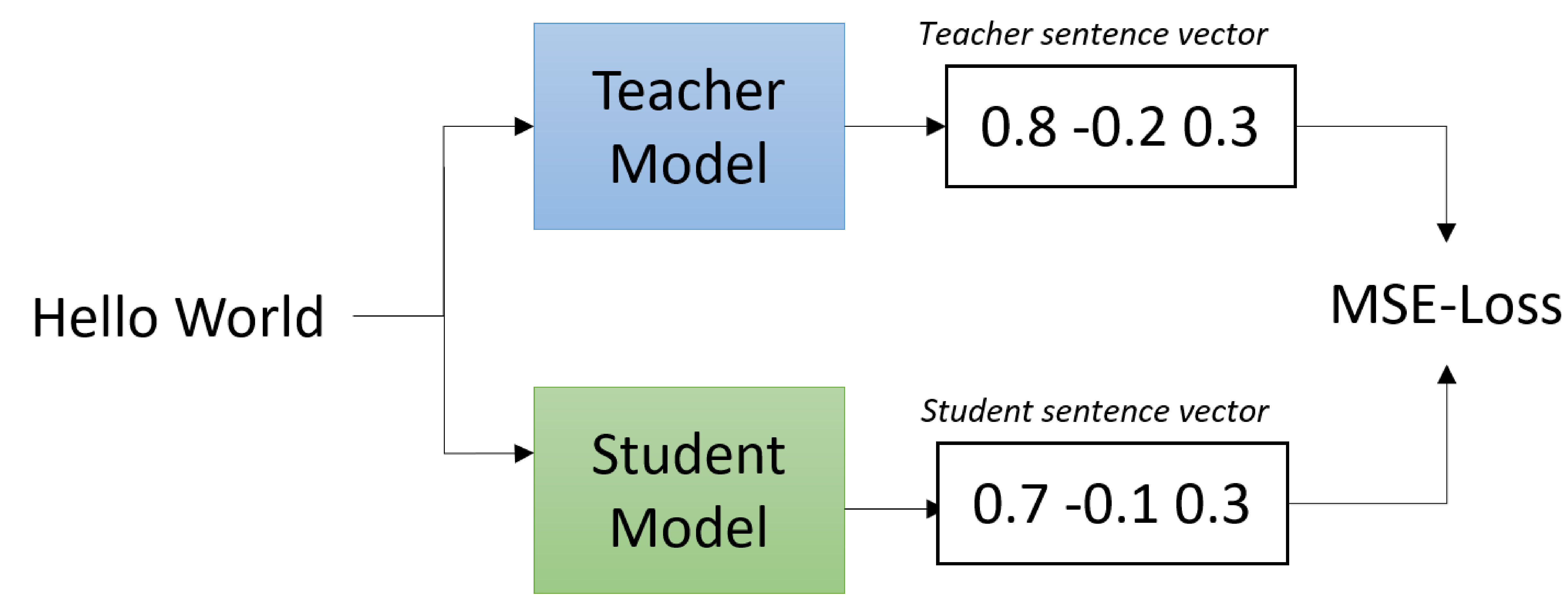
MODEL DISTILLATION : INCREASING SIZE & SPEED

2 WAYS OF MODEL DISTILLATION

- **DIMENSIONALITY REDUCTION**
- **PARAMETER QUANTISATION**
- **REDUCING NUMBER OF LAYERS**

IDEA IS TO EITHER REDUCE THE SPACE USED BY THE VECTOR EMBEDDINGS OR THE PROCESSING TIME

TEACHER MODEL TRANSFERS ITS KNOWLEDGE INTO THE STUDENT MODEL



Layers	STSbenchmark Performance	Performance Decrease	Speed (Sent. / Sec. on V100-GPU)
teacher: 12	85.44	-	2300
8	85.54	+0.1%	3200
6	85.23	-0.2%	4000
4	84.92	-0.6%	5300
3	84.39	-1.2%	6500
2	83.32	-2.5%	7700
1	80.86	-5.4%	9200

HTTPS://GITHUB.COM/INSIGHTBUILDER

MODEL RANKINGS: SIMILARITY & MORE

Model	Performance (14 sentence similarity tasks)
microsoft/mpnet-base	60.99
nghuyong/ernie-2.0-en	60.73
microsof/deberta-base	60.21
roberta-base	59.63
t5-base	59.21
bert-base-uncased	59.17
distilbert-base-uncased	59.03
nreimers/TinyBERT_L-6_H-768_v2	58.27
google/t5-v1_1-base	57.63

MTEB Leaderboard - a Hugging Face Space by mteb

Discover amazing ML apps made by the community

🤗 huggingface

Model	Embedding Dimensions	Sequence Length	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)	STS Average (10 datasets)	Summarization Average (1 dataset)	
1	e5-large-v2	1024	512	62.25	75.24	44.49	86.03	56.61	50.56	82.05	30.19
2	instructor-xl	768	512	61.79	73.12	44.74	86.62	57.29	49.26	83.06	32.32
3	instructor-large	768	512	61.59	73.86	45.29	85.89	57.54	47.57	83.15	31.84
4	e5-base-v2	768	512	61.5	73.84	43.8	85.73	55.91	50.29	81.05	30.28

HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

REMEMBER TO PRACTICE WITH EXAMPLES

LIKE

SHARE

SUBSCRIBE

HTTPS://GITHUB.COM/INSIGHTBUILDER