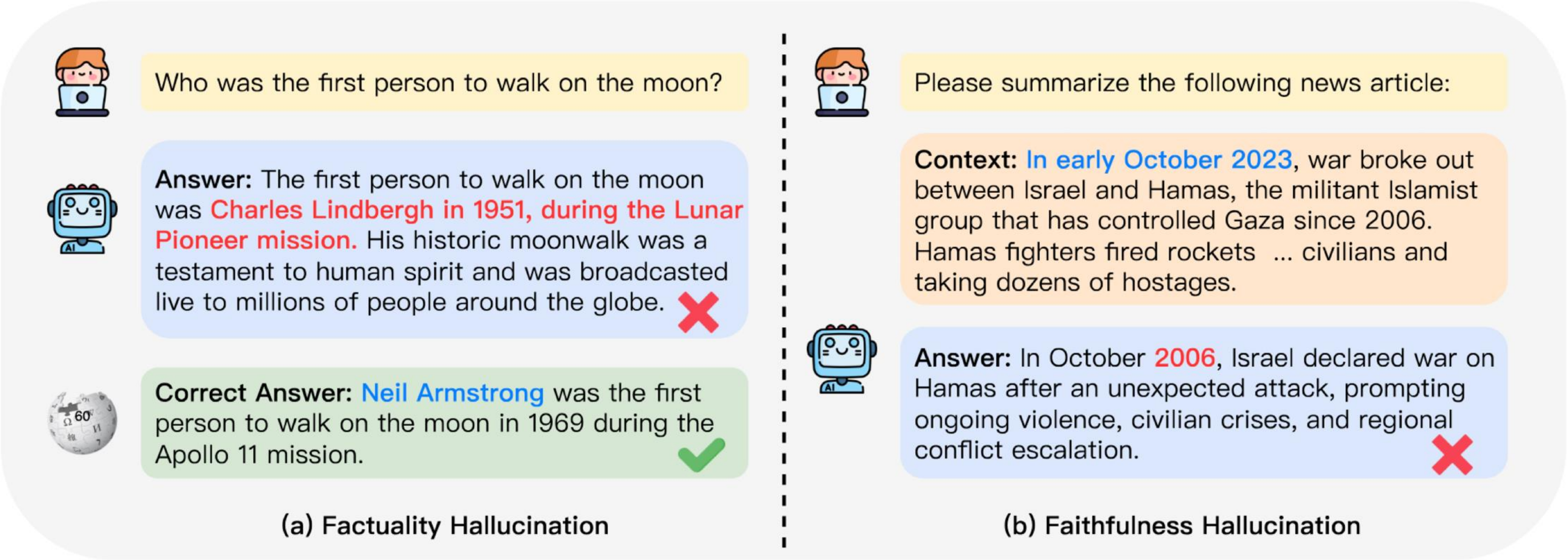
Natural Language Processing

Retrieval-Augmented Generation

Hallucination of LLM

- It is discovered that NLG models often generate text that is nonsensical, or unfaithful to the provided input. Such undesirable generation is referred to Hallucination (Ji et al., 2023).



Ji et al. "Survey of hallucination in natural language generation." ACM Computing Surveys 55.12 (2023): 1-38.
Figure source: Munkhdalai, Tsendsuren, Manaal Faruqui, and Siddharth Gopal. "Leave no context behind: Efficient infinite context transformers with infini-attention." arXiv preprint arXiv:2404.07143 (2024).

Solutions to Mitigating Hallucinations

- Chain-of-Thought Prompting (CoT)
- Retrieval-Augmented Generation (RAG)
- …

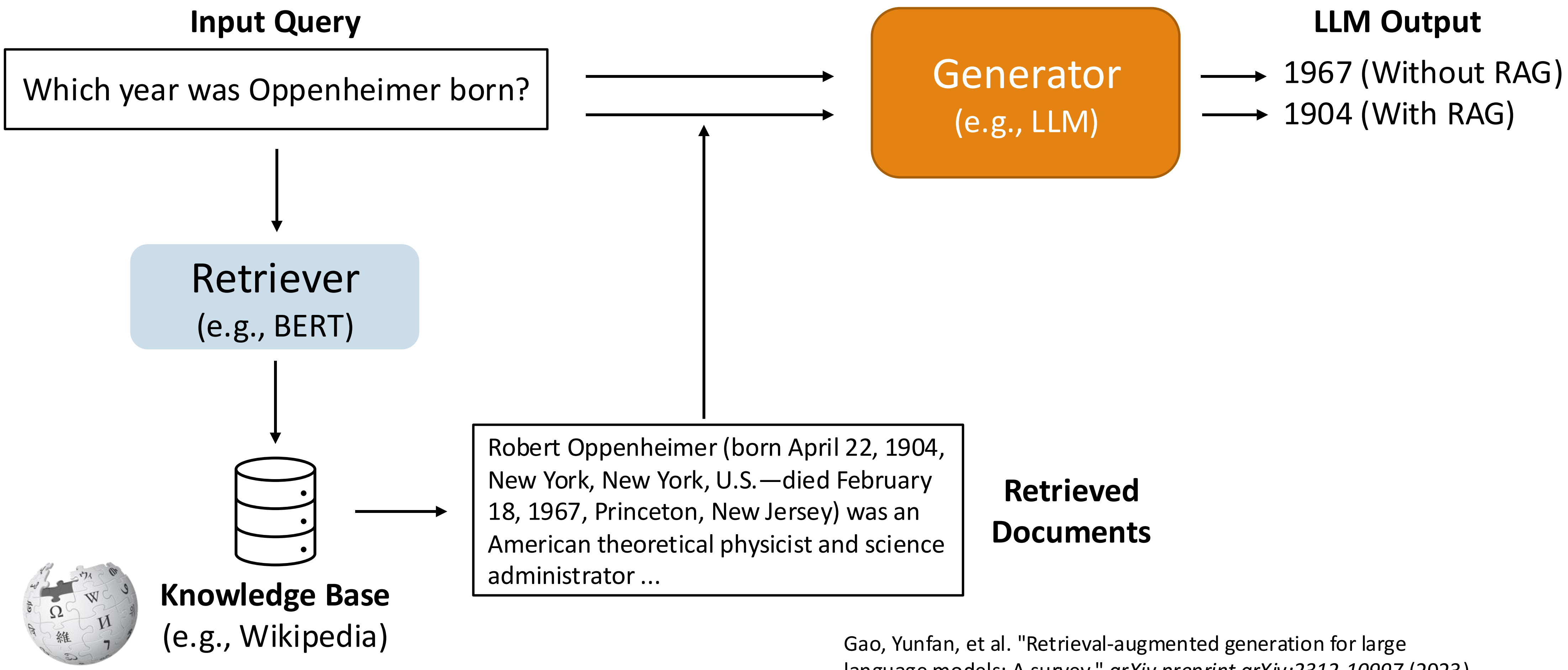
Retrieval-Augmented Generation (RAG)

Information Retrieval

- Retrieval: get relevant information from a pool (like a search engine)

- Retrieval-Augmented Generation (RAG): Perform generation with additional retrieved information

Retrieval-Augmented Generation (RAG)



Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* (2023).

Why do we need RAG?

- LLMs have profound parameterized knowledge that makes them useful in responding to general prompts.
- However, LLMs are error-prone due to a lack of domain knowledge or outdated information.
- Standalone LLMs do not serve users who want a deeper dive into a current or more specific topic.

https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

Retrievers for RAG

- A retriever is aimed at searching relevant documents based on an input query.
- A retriever plays an important role in enhancing the performance of an LLM. Therefore, a good retriever is needed.
- Usually, a retriever produces outputs by computing similarities between query embeddings and document embeddings, which come from other encoder models or the retriever itself.

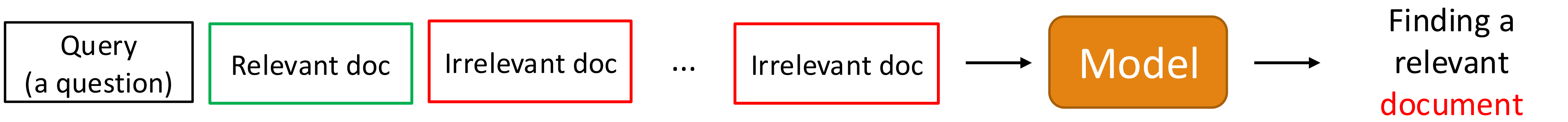
Embedding Types for Retrieval

- Sparse Embeddings (sparse vectors)
  - E.g., TF-IDF, BM25
- Dense Embeddings (dense vectors)
  - E.g., BERT, Sentence-BERT, DPR (Dense Passage Retrieval)

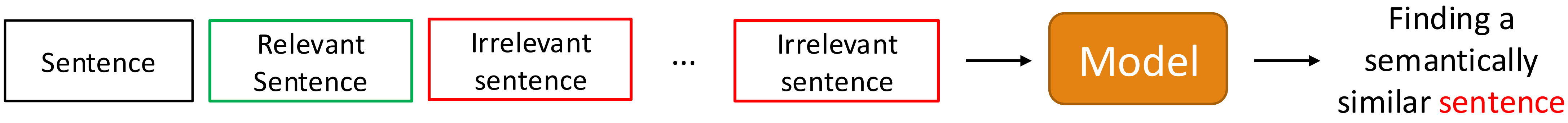
Retrievers for RAG

- Task-oriented training:
  - Retrieval for open-domain question answering (ODQA)
    - Usually used in RAG because it is more common to search relevant documents.
  - Sentence embeddings for semantic similarity tasks

Retrieval for open-domain question answering



Sentence embeddings for semantic similarity tasks



- Both approaches are suitable for retrieval (depends on your task for an LLM).

Sentence embeddings for semantic similarity tasks

Sparse Vectors

- In information retrieval, sparse vectors are vectors with most elements set to zero.
- The advantage of sparse vectors is computational efficiency because most elements are zero and can be ignored.

For an example:
We have a small vocabulary with five words: ["cat", "dog", "fish", "bird", "snake"].
We have a document that only contains the words "cat" and "dog".
The sparse vector for this document would look like this: [1, 1, 0, 0, 0]

The elements represented to "fish", "bird", and "snake" are 0s

Sparse Embeddings: Bag-of-words

```
texts = [
    "This is a book",
    "These are pens and my pen is here"
]
```

- The Bag-of-words approach creates document embeddings.
- The embedding size is equal to the vocabulary size.
- Each value of an embedding is based on frequency counts.

Transform via frequency

Vocabulary size

	a	and	are	book	here	is	my	pen	these	this
sent_0	1	0	0	1	0	1	0	0	0	1
sent_1	0	1	1	0	1	1	1	2	1	0

Since the outputs contain many zeros, this approach is called a sparse embedding method.

Sparse Embeddings: TF-IDF

- TF (Term Frequency)
- IDF (Inverse Document Frequency)

```
texts = [
    "This is a book",
    "These are pens and my pen is here"
]
```

Transform via TF-IDF

- The **TF-IDF** approach also creates document embeddings.
- The embedding size is equal to the vocabulary size.
- Each value of an embedding is based on **TF x IDF**.

Vocabulary size

	a	and	are	book	here	is	my	pen	these	this
sent_0	0.534046	0.000000	0.000000	0.534046	0.000000	0.379978	0.000000	0.000000	0.000000	0.534046
sent_1	0.000000	0.324336	0.324336	0.000000	0.324336	0.230768	0.324336	0.648673	0.324336	0.000000

Since the outputs contain many zeros, this approach is called a sparse embedding method.

https://github.com/tsmatz/nlp-tutorials/blob/master/01_sparse_vector.ipynb

TF-IDF

The mathematical representation of TF-IDF:

$$TF - IDF = TF \times IDF$$ where $$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad IDF_i = lg\frac{|D|}{|\{j : t_i \in d_j\}|}$$

- Where $n_{i,j}$ is the i-th word in j-th text in the dataset.

TF (Term Frequency)

- Represents the "frequency" of a term appearing in a text.

IDF (Inverse Document Frequency)

- Aims for terms to have higher specificity, meaning the fewer texts in the dataset contain the term, the better.

BM25 (an improved version of TF-IDF)

$$score = \frac{(k_1 + 1)TF}{TF + k_1 * (1 - b + b * \frac{|D|}{avgD})} * IDF,\ b \in [0,1]$$

k_1 : A term frequency saturation hyper-parameter. For best performance, the value of k_1 should be between 0 and 3.

- This reduces the effect of high-frequency terms so that they don't overpower the score excessively.

b : A document length normalization parameter, this hyper-parameter controls the influence of sequence length.

- This allows shorter and longer documents to compete more equally in retrieval relevance.

Robertson, Stephen, and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." *Foundations and Trends® in Information Retrieval* 3.4 (2009): 333-389.

Dense Vectors

- In NLP, dense vectors comprise compact numerical values representing semantic features of text.
- “Dense” is concept contrary to “sparse.”
- Word2vec is also an approach for creating dense embeddings.

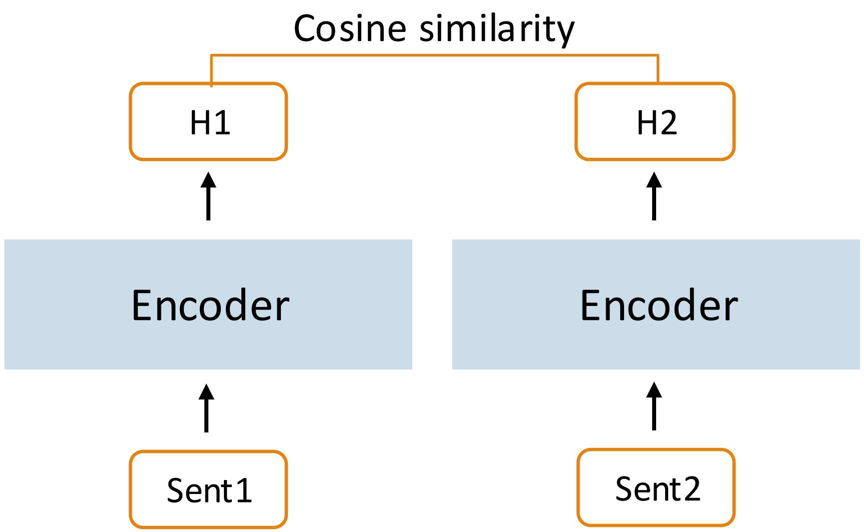
Sparse vector

Dense vector

Approach for Dense Vectors: Dual Encoder

- Also called **bi-encoder**, **Siamese network.**
- Structure:
  - Two identical/similar encoders
  - Processes two inputs independently
  - Outputs separate vectors for each input
- Tasks:

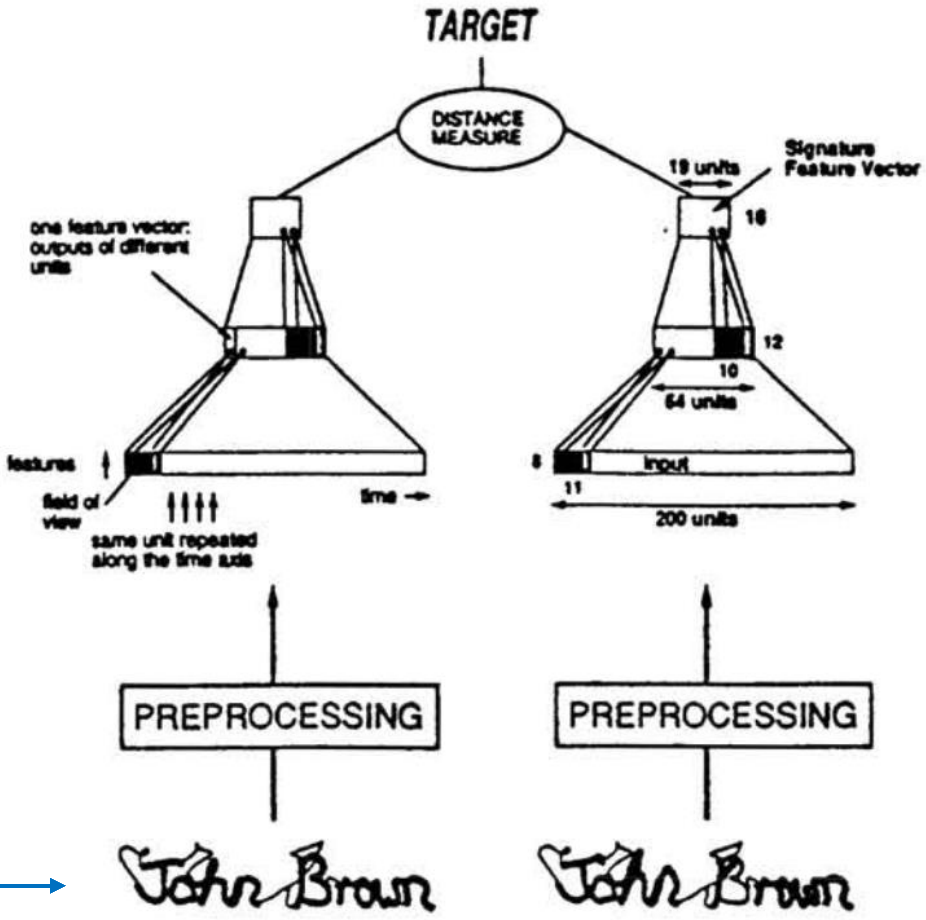
Task	Inputs
Information Retrieval	Document and query
Semantic similarity (or any sentence pair classification)	Two sentences



The First Siamese Network

- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a" siamese" time delay neural network. NeurIPS.
- Siamese" neural network consists of **two identical** sub-networks joined at their outputs.

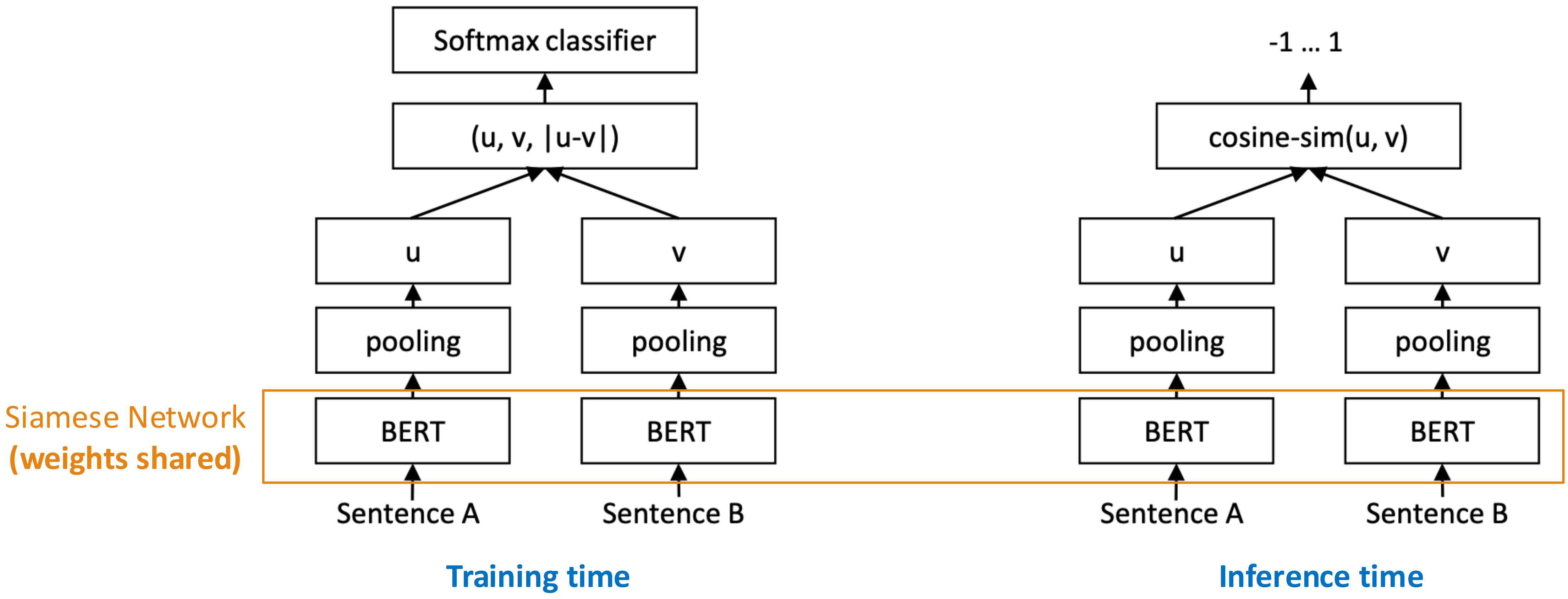
Signature Verification



(Figure source: Bromley et al., 1993)

Sentence-BERT

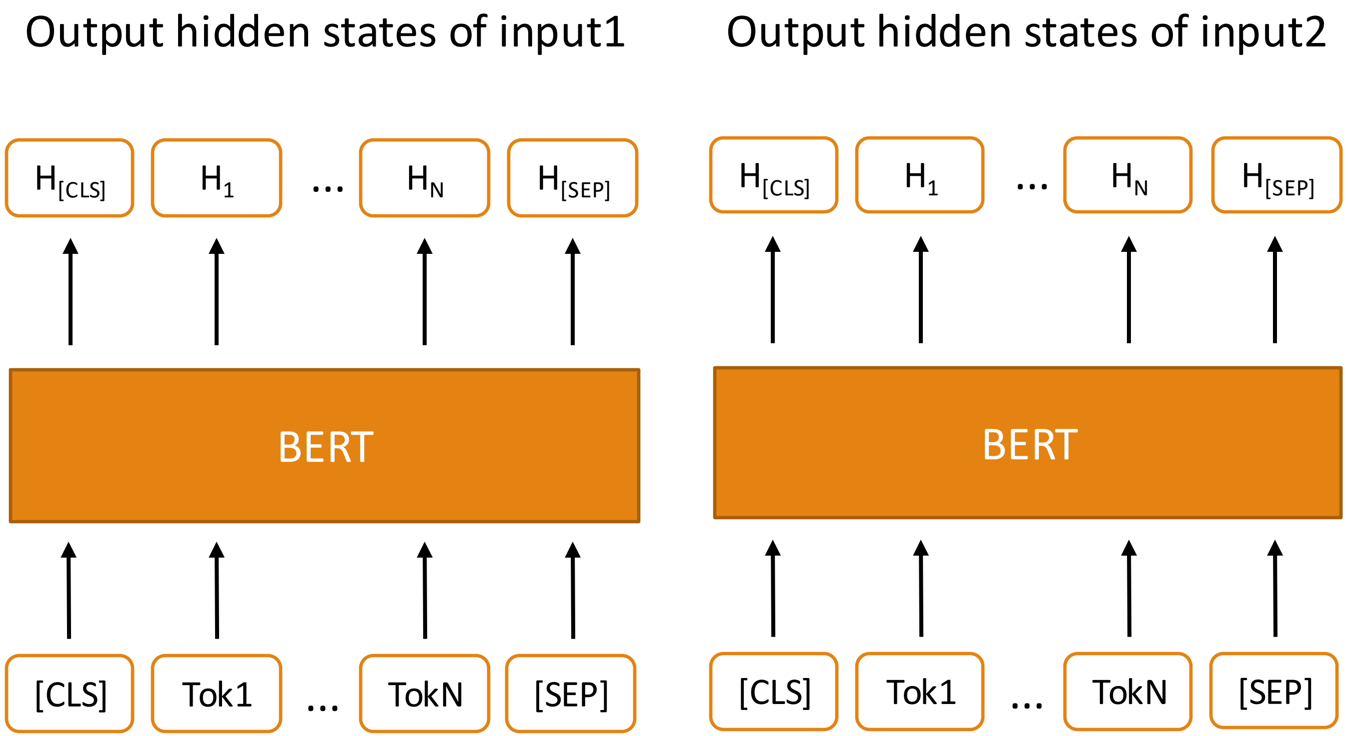
```
# Pseudo code for Dual Encoder
query_vector = encoder(query)            # [0.1, 0.2, 0.3]
document_vector = encoder(document)      # [0.2, 0.2, 0.4]
similarity = cosine_similarity(query_vector, document_vector)
```



Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *(EMNLP-IJCNLP 2019)*

Why does Sentence-BERT need pooling?

- BERT produces embeddings (hidden states from the final layer) for each token.
- We need a single fixed-size vector for the entire sentence.



Pooling of Sentence-BERT

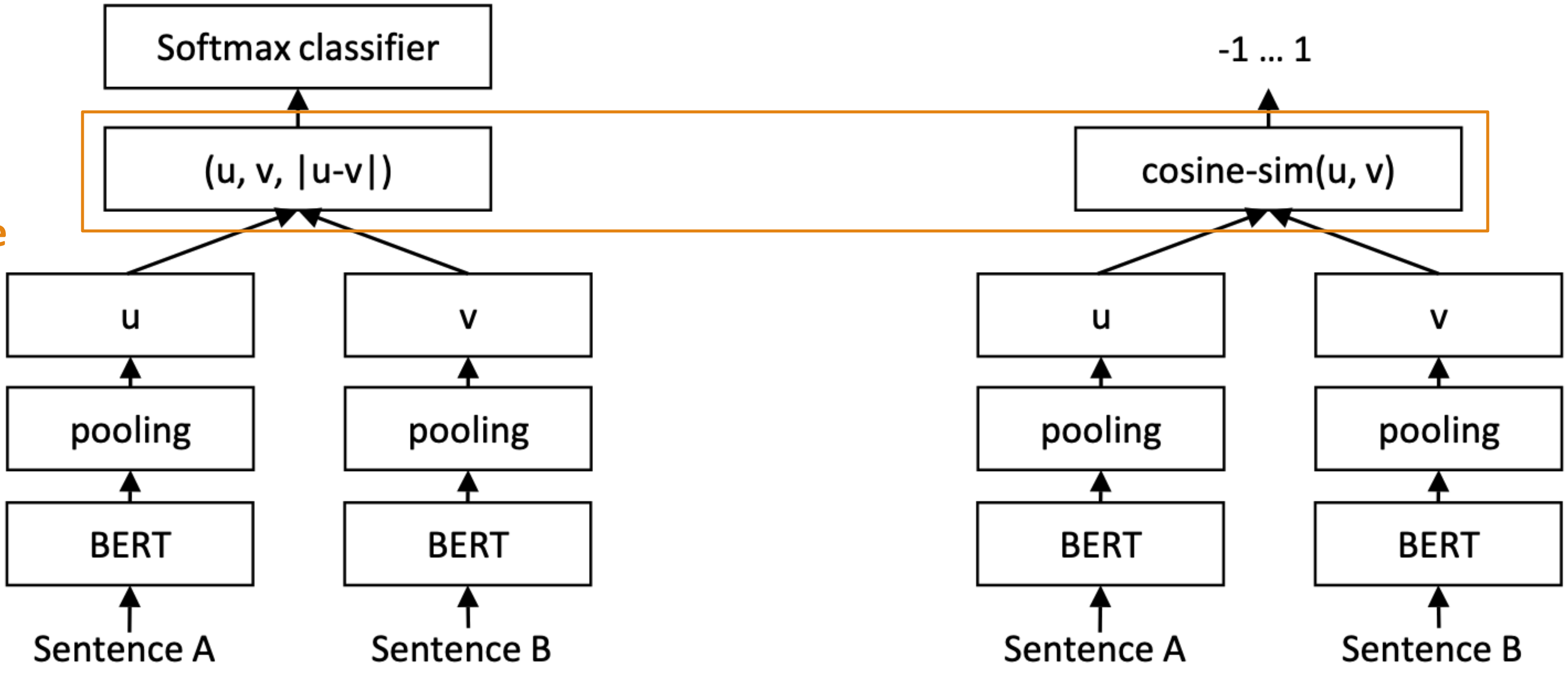
We need a single fixed-size vector for the entire sentence.

- **CLS: Use the [CLS] token**
  - This is the default setting in original BERT.
- **MEAN: the mean of all output vectors**
  - Averages all token embeddings.
- **MAX: max-over-time of the output vectors**
  - Takes maximum value across each dimension.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *(EMNLP-IJCNLP 2019)*

Sentence-BERT(Dual encoder)

Output Concatenation with element-wise difference



Training time **Inference time**

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *(EMNLP-IJCNLP 2019)*

Performance comparison of Pooling and Concatenation

- Indeed, [CLS] token can directly be used for representing the entire sentence.
- But pooling may bring better performance.

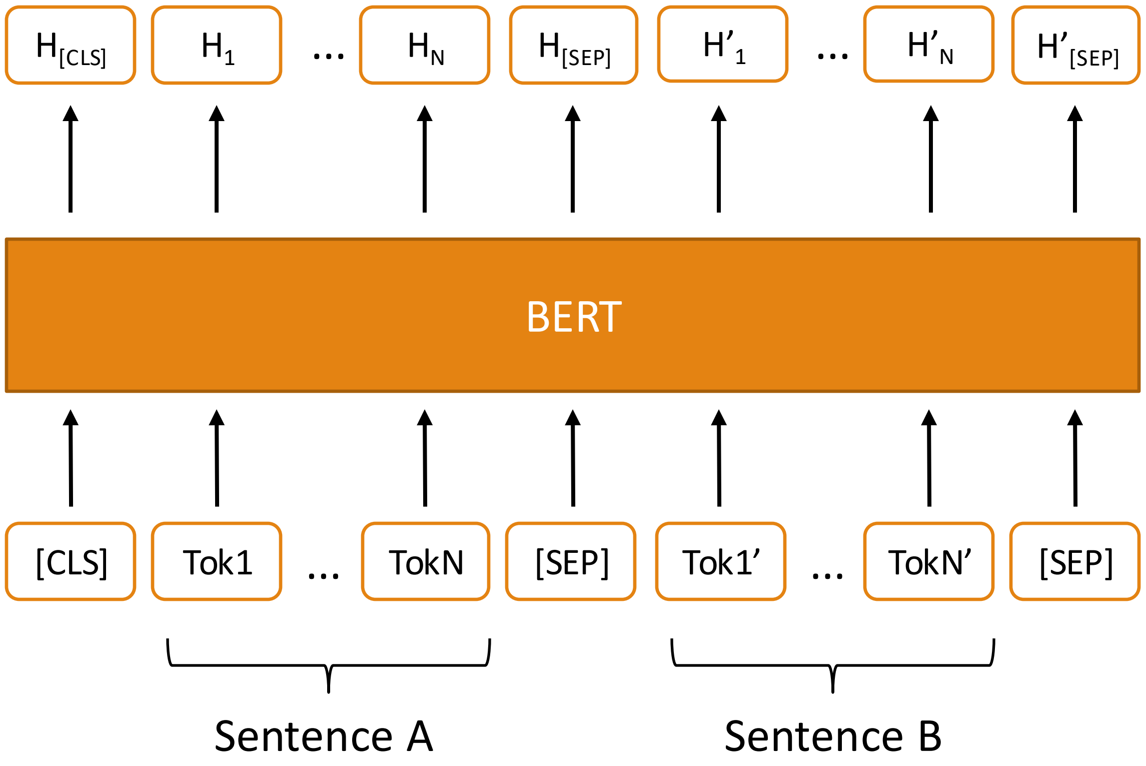
- **Experiment** shows using element-wise difference is better than the other settings.
- Note that the concatenation mode is only used for training.

	NLI	STSb
Pooling Strategy		
MEAN	**80.78**	**87.44**
MAX	79.07	69.92
CLS	79.80	86.62
Concatenation		
(u, v)	66.04	-
$(\lvert u - v \rvert)$	69.78	-
$(u * v)$	70.54	-
$(\lvert u - v \rvert, u * v)$	78.37	-
$(u, v, u * v)$	77.44	-
$(u, v, \lvert u - v \rvert)$	**80.78**	-
$(u, v, \lvert u - v \rvert, u * v)$	80.44	-

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *(EMNLP-IJCNLP 2019)*

BERT as a Cross Encoder

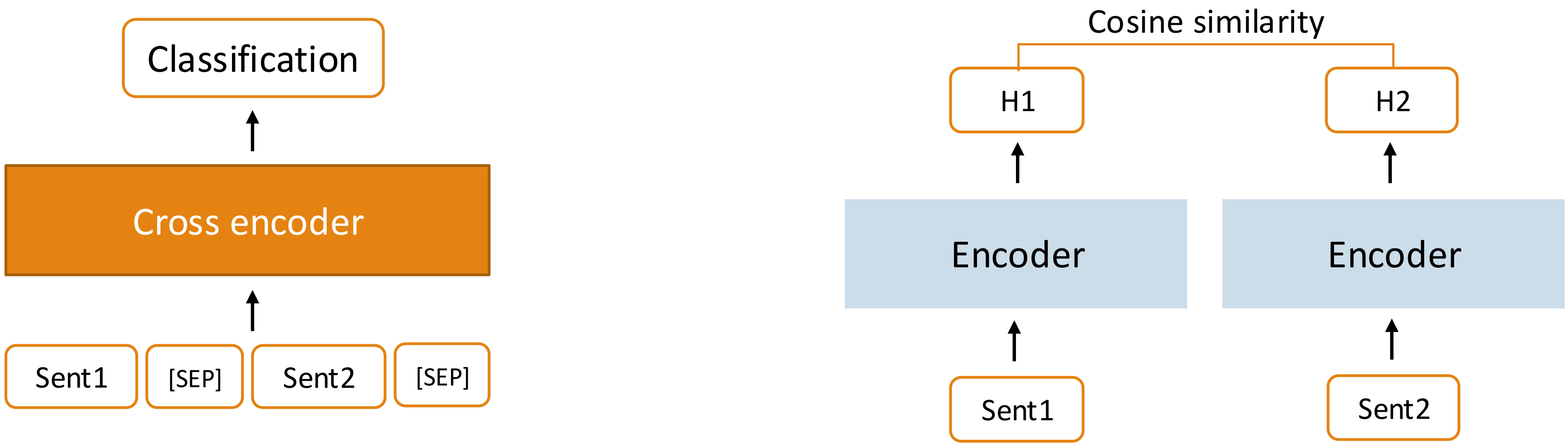
- For a cross encoder, representations of two input sentences are attended with each other.
- The hidden state of [CLS] represents the relationship between the two input sentences.



Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

Computation time for Bi-encoders and Cross encoders

- For 10,000 sentence pairs:
  - Cross encoders: $n \cdot (n-1)/2$ = 49,995,000 inference times
  - Bi-encoders: 10,000 * 2 inference times (can be parallel) with cosine similarity calculation



Performance comparison for Bi-encoders and Cross encoders

Model	Spearman
Not trained for STS	
Avg. GloVe embeddings	58.02
Avg. BERT embeddings	46.35
InferSent - GloVe	68.03
Universal Sentence Encoder	74.92
SBERT-NLI-base	77.03
SBERT-NLI-large	79.23
Trained on STS benchmark dataset	
BERT-STSb-base	84.30 ± 0.76
SBERT-STSb-base	84.67 ± 0.19
SRoBERTa-STSb-base	$\mathbf{84.92} \pm 0.34$
BERT-STSb-large	$\mathbf{85.64} \pm 0.81$
SBERT-STSb-large	84.45 ± 0.43
SRoBERTa-STSb-large	85.02 ± 0.76
Trained on NLI data + STS benchmark data	
BERT-NLI-STSb-base	$\mathbf{88.33} \pm 0.19$
SBERT-NLI-STSb-base	85.35 ± 0.17
SRoBERTa-NLI-STSb-base	84.79 ± 0.38
BERT-NLI-STSb-large	$\mathbf{88.77} \pm 0.46$
SBERT-NLI-STSb-large	86.10 ± 0.13
SRoBERTa-NLI-STSb-large	86.15 ± 0.35

BERT: Cross encoder

SBERT / SRoBERTa: Bi-encoders

- Generally, the difference in performance between bi-encoders and cross encoders is not large.
- However, bi-encoders are much faster with respect to computation time.
  - If >1M documents in a database, the difference in computation time will be extremely huge.

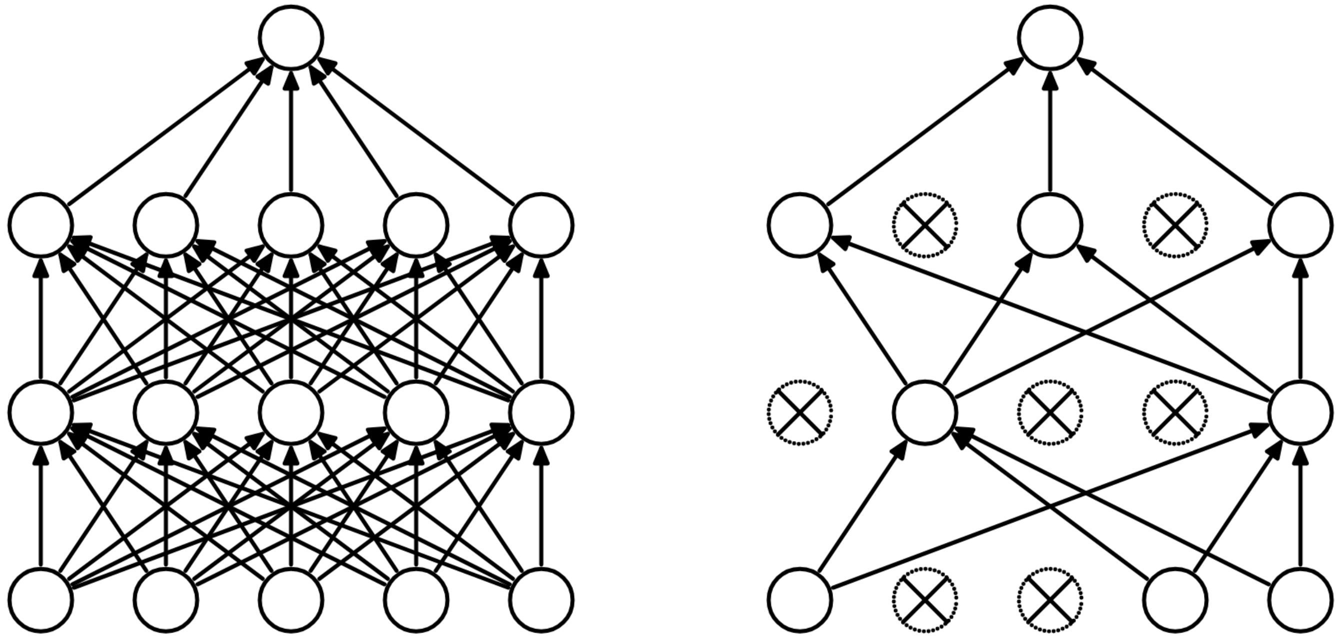
SimCSE (Dual encoder)

- SimCSE: a simple contrastive sentence embedding framework
- Both unsupervised and supervised training approaches were proposed in SimCSE:
  - **Unsupervised** training of SimCSE
    - Relying on **Dropout**
  - **Supervised** training of SimCSE
    - Relying on **labels in a dataset**

Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings." EMNLP 2021.

Dropout

- Dropout randomly drop units (along with their connections) from the neural network during training. This approach usually brings regularization and reduces overfitting.

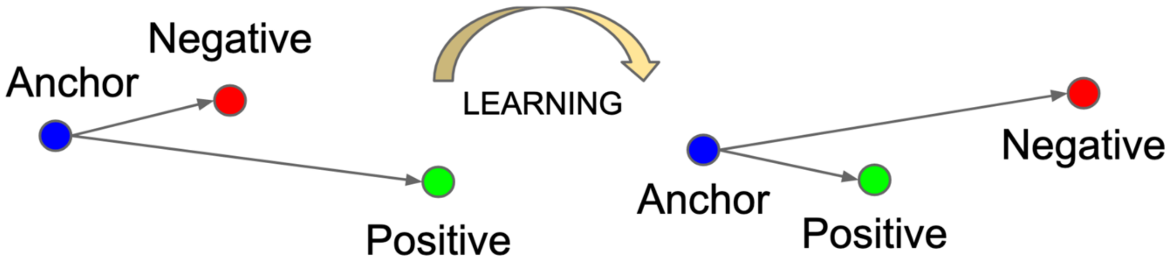


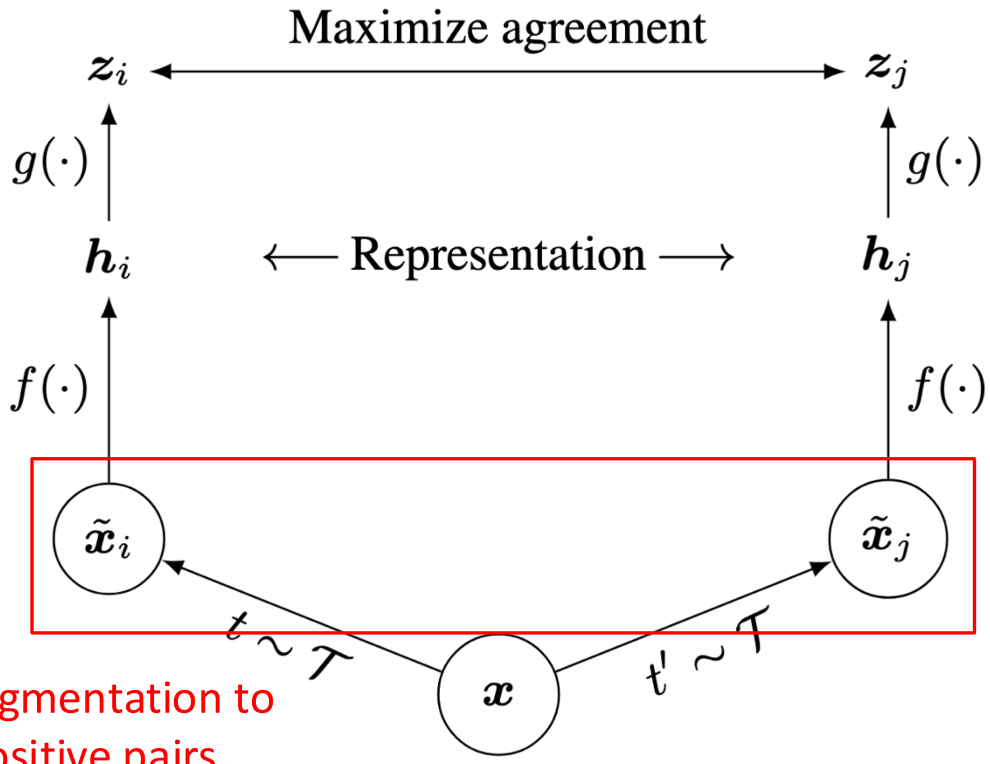
(a) Standard Neural Net (b) After applying dropout.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.

Contrastive Learning

f (·): encoder network
g(·): projection head
z: output logits



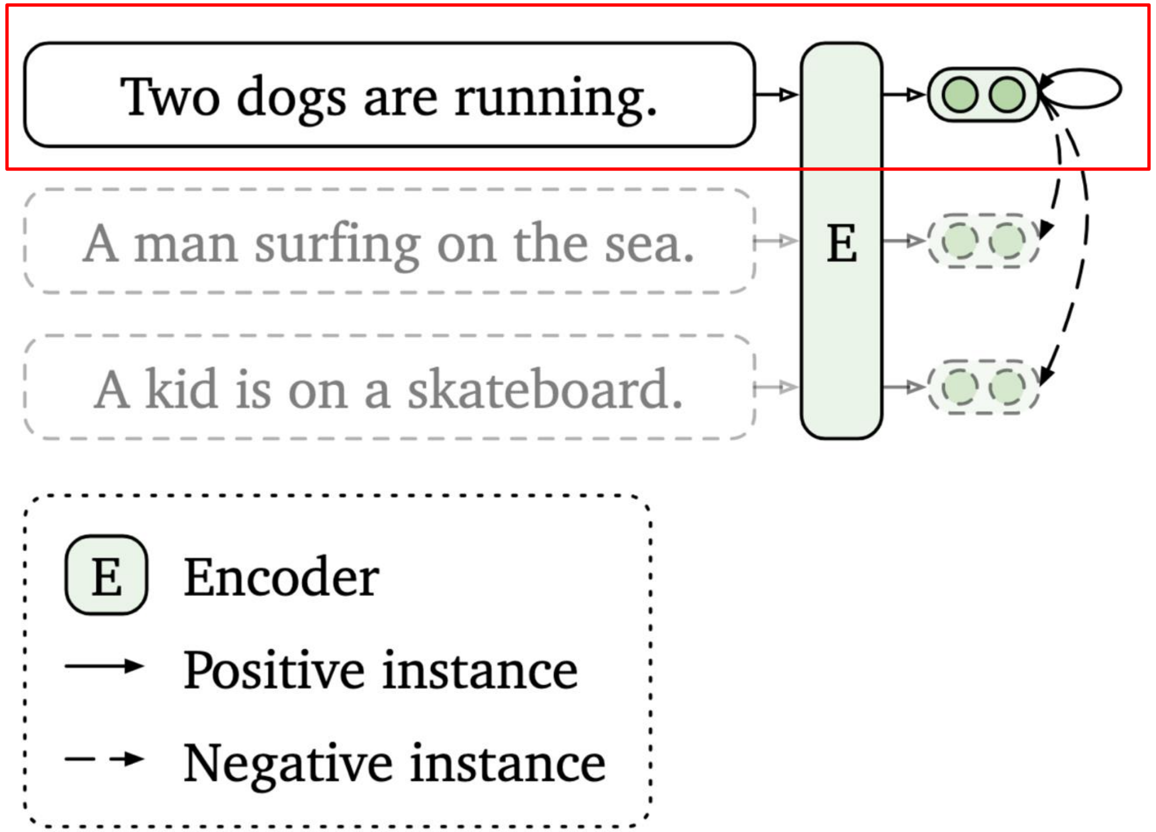


Use data augmentation to create positive pairs

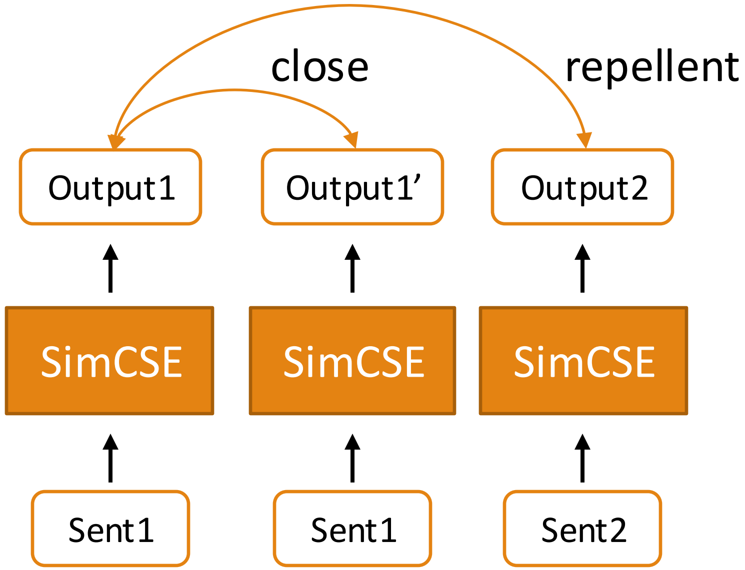
Left Figure source: Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." CVPR 2015.

Right Figure source: Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." ICLR 2020.

Unsupervised training of SimCSE



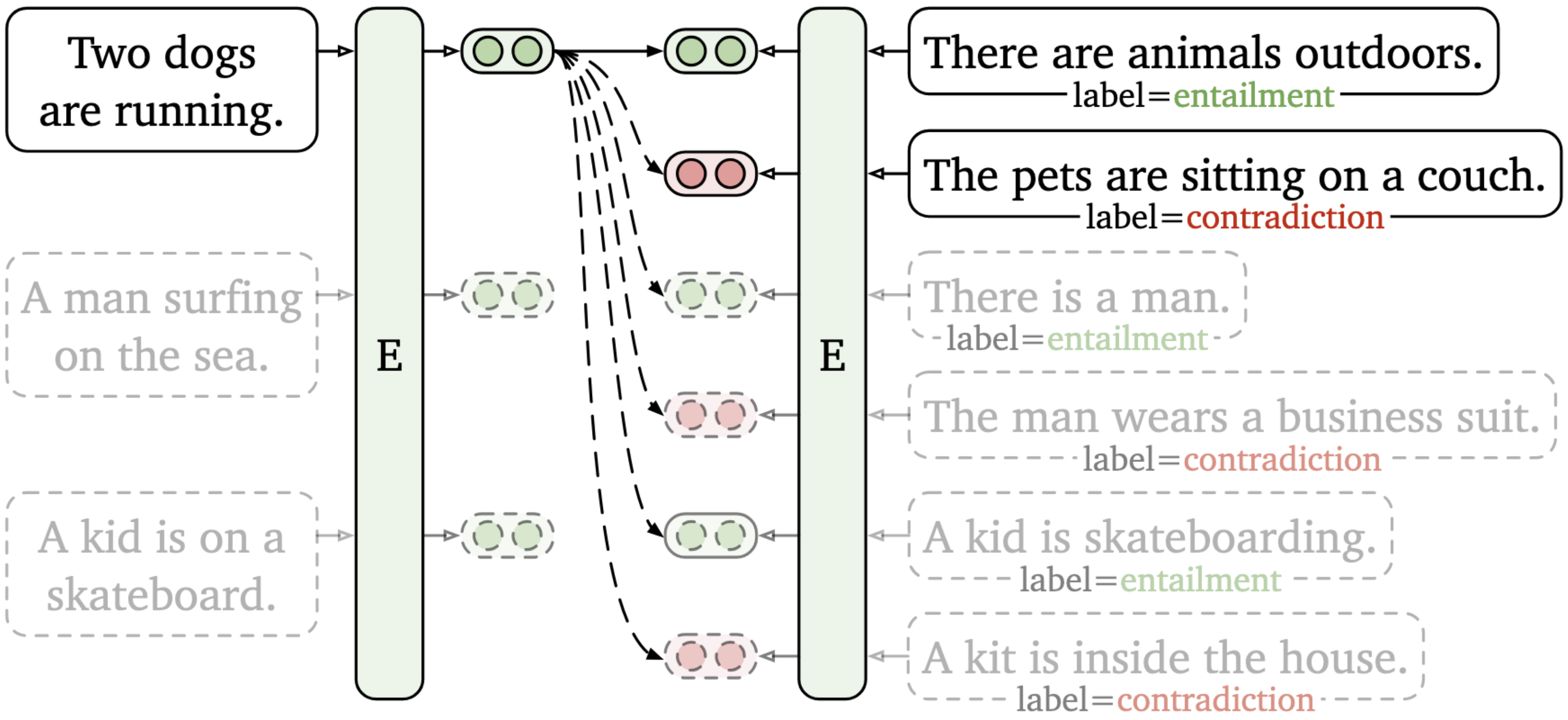
Input a sentence twice with a dropout mask
(Dropout as data augmentation)



Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings." EMNLP 2021.

Supervised training of SimCSE

- Supervised training of SimCSE relies on labels in a dataset to define positives and negatives.



Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings." EMNLP 2021.

SimCSE outperforms Sentence-BERT

- SimCSE outperforms Sentence-BERT on semantic similarity tasks.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
				Supervised models				
$\text{SRoBERTa}_{\texttt{base}}$ ♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
$\text{SRoBERTa}_{\texttt{base}}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
$*$ $\text{SimCSE-RoBERTa}_{\texttt{base}}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
$*$ $\text{SimCSE-RoBERTa}_{\texttt{large}}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

Retrieval for open-domain question answering (ODQA)

Open-domain question answering (ODQA)

- Given a question x such as "What is the currency of the UK?", a model must output the correct answer string y, "pound".
- The "open" part of ODQA refers to the fact that the model does not receive a pre-identified document that is known to contain the answer.
- ODQA is like Reading comprehension (RC) tasks, such as SQuAD, but no relevant articles provided.

(Example of SQuAD)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

Guu, Kelvin, et al. "Retrieval augmented language model pre-training." ICML 2020.
Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP 2016.

Dense Passage Retrieval (DPR)

- Dense retrieval focuses on **semantic similarity**
- Passages and questions are embedded into dense vectors
- Dense vectors enable better matching for related words or phrases

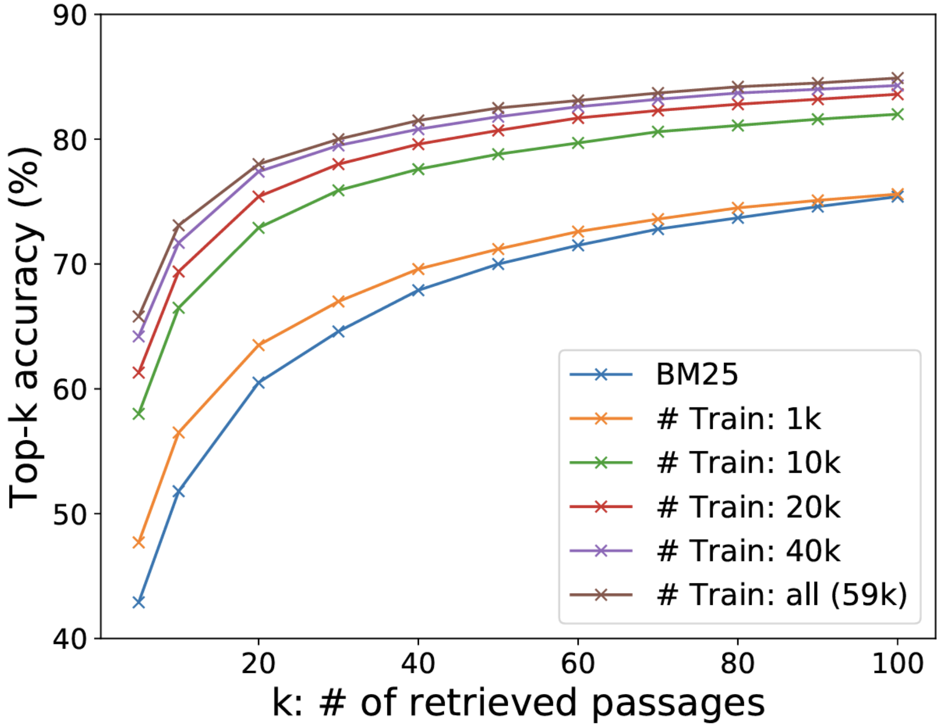
(e.g., “the body of water” matched with “sea”)

Question	Passage received by BM25	Passage retrieved by DPR
What is the body of water between England and Ireland?	Title:British Cycling ...**England** is not recognised as a region by the UCI, and there is no English cycling team outside the Commonwealth Games. For those occasions, British Cycling selects and supports the **England** team. Cycling is represented on the Isle of Man by the Isle of Man Cycling Association. Cycling in Northern **Ireland** is organised under Cycling Ulster, part of the all-Ireland governing **body** Cycling **Ireland**. Until 2006, a rival governing **body** existed, ...	Title: Irish Sea ...Annual traffic between Great Britain and **Ireland** amounts to over 12 million passengers and of traded goods. **The Irish Sea** is connected to the North Atlantic at both its northern and southern ends. To the north, the connection is through the North Channel between Scotland and Northern **Ireland** and the Malin Sea. The southern end is linked to the Atlantic through the St George’s Channel between Ireland and Pembrokeshire, and the Celtic Sea. ...

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval (DPR)

- Outperforms **BM25** using only 1000 training data!



Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval (DPR)

- After training, two **BERT-based encoders** can **independently** encode question (**q**) and passage (**p**) into dense vectors.

- **Similarity** between question and passage = **dot product** between their embeddings

$$\mathrm{sim}(q,p) = E_Q(q)^\intercal E_P(p).$$

q : question text
p : passage text
E_Q : BERT model that outputs question representation
E_p : BERT model that outputs passage representation

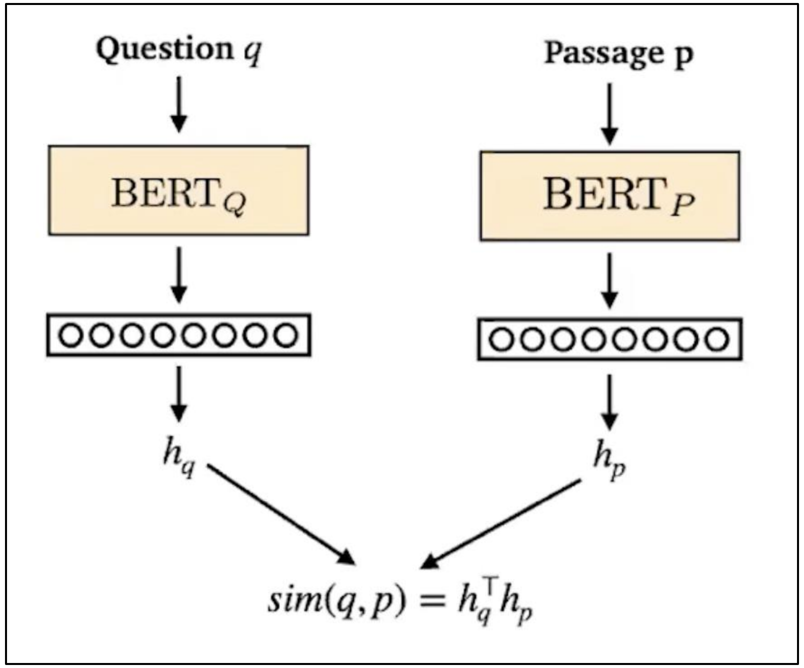


Image source: Stanford CS224N Lecture 12 - Question Answering

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval (DPR)

Training the encoders

- Goal: **Relevant** pairs of questions and passages will have **smaller distance** than the irrelevant ones

- Training data

m training instances

$$\mathcal{D} = \{\langle \boxed{q_i}, \boxed{p_i^+}, \boxed{p_{i,1}^-, \cdots, p_{i,n}^-} \rangle\}_{i=1}^{\boxed{m}}$$

Question

Relevant Passage

n Irrelevant Passages

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval (DPR)

Training the encoders

- Base model: bert-base-uncased

- Loss function: Negative log-likelihood of the positive passage

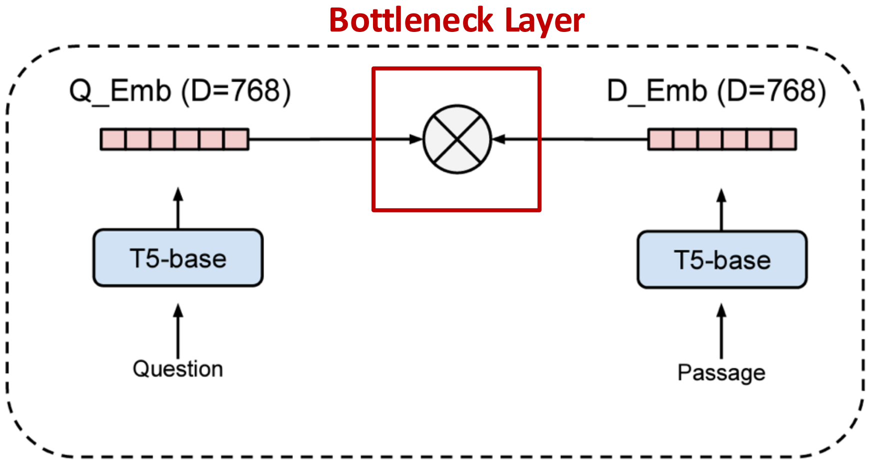
$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}.$$

  - **Maximize** the similarity between $\mathbf{q}_i$ and $\mathbf{p}_i^+$
  - **Minimize** the similarity between non-relevant pairs ($\mathbf{q}_i$ and $\mathbf{p}_{i,j}^-$)

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

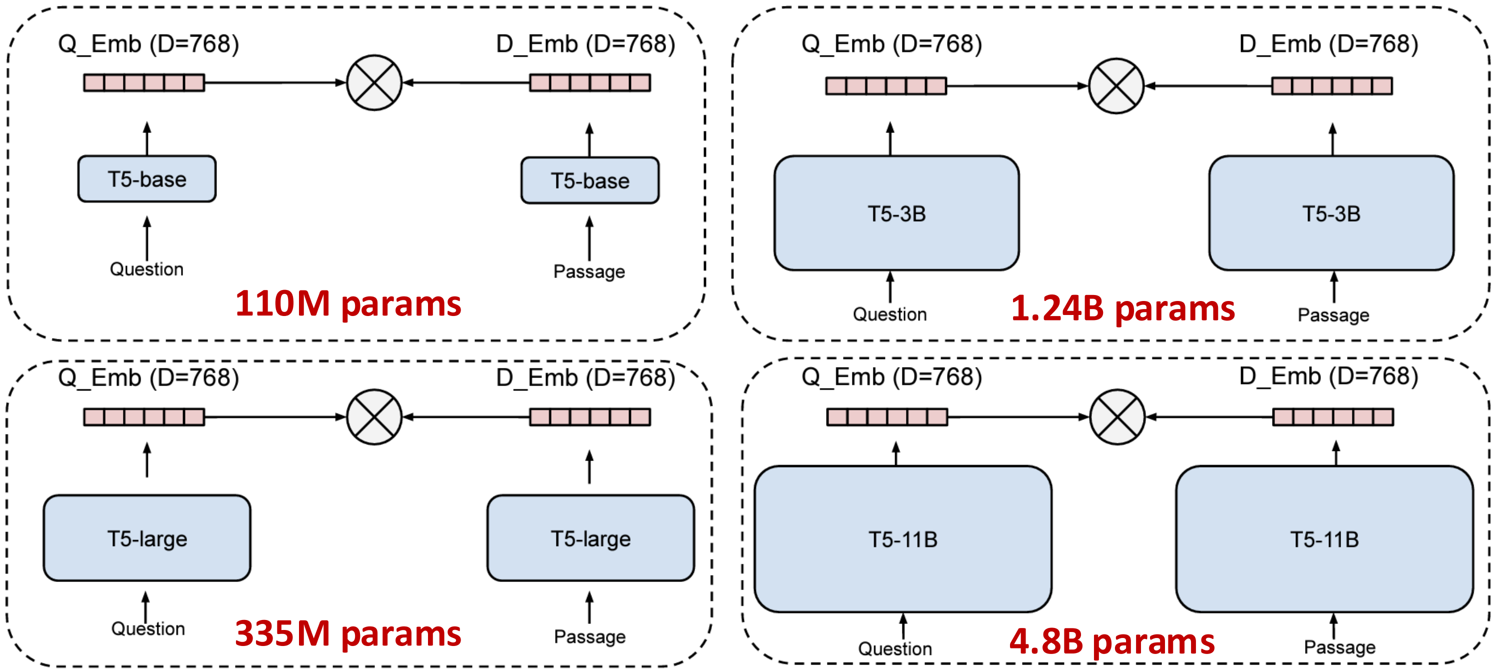
Limitations of Dual Encoders

- Often fail to generalize to other domains for retrieval tasks
- **Bottleneck layer** of dual encoders (simple dot-product or cosine similarity) *might not be powerful enough* to **capture semantic relevance**?



Ni et al., 2021. Large Dual Encoders Are Generalizable Retrievers

Generalizable T5-based Retriever (GTR)

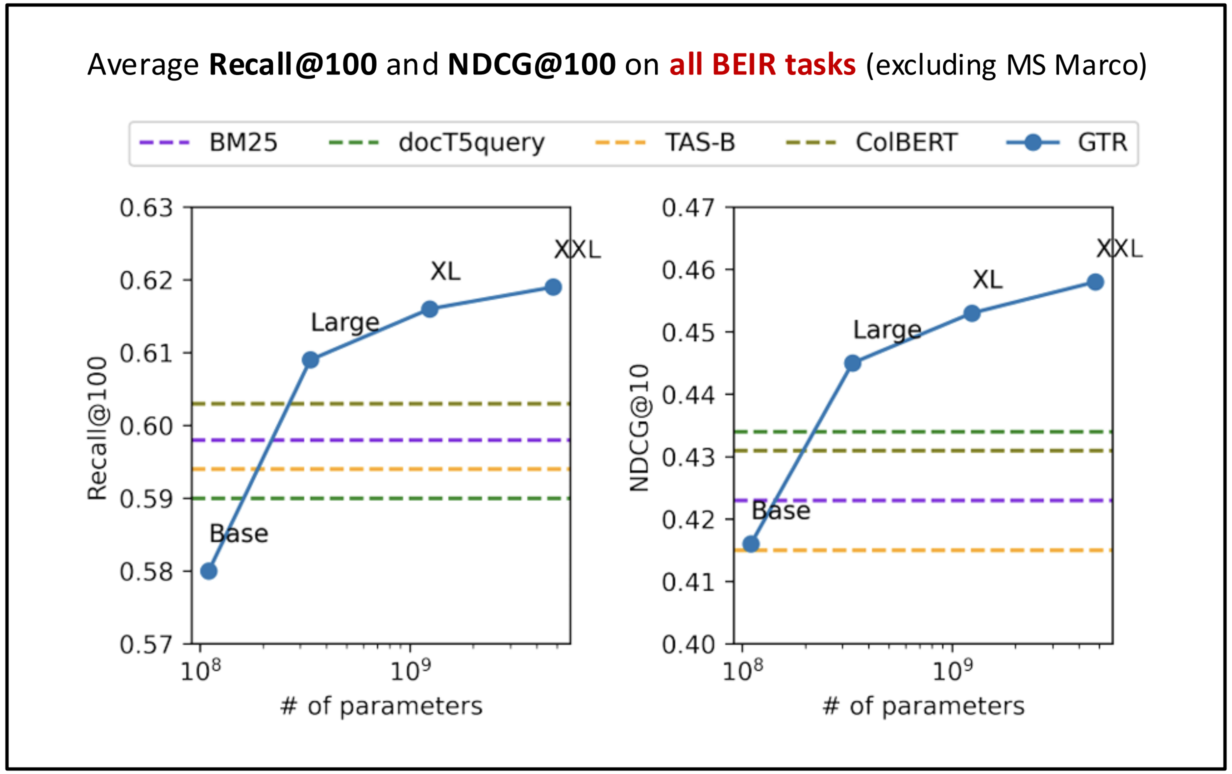


Can **scaling up** dual encoder model size *improve the retrieval performance*, while keeping the bottleneck layers **fixed**?

Ni et al., 2021. Large Dual Encoders Are Generalizable Retrievers

Generalizable T5-based Retriever (GTR)

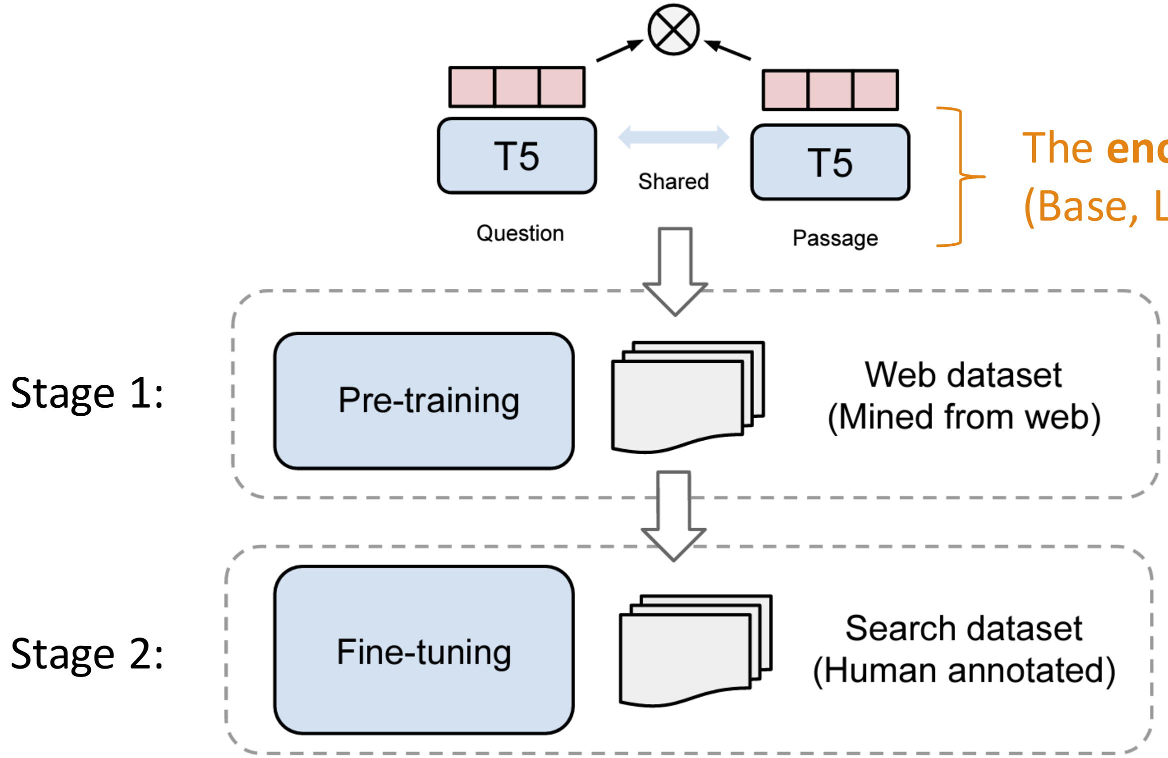
- Scaling up *consistently improves* dual encoders' **out-of-domain** performance.



Ni et al., 2021. Large Dual Encoders Are Generalizable Retrievers

Generalizable T5-based Retriever (GTR)

Multi-stage training for GTR



The **encoder** part from pretrained T5 models (Base, Large, XL, XXL)

2 billion **QA pairs** from:
- Reddit
- Stack-Overflow

Dataset:
- MS Marco
- NaturalQuestions

Ni et al., 2021. Large Dual Encoders Are Generalizable Retrievers

Generalizable T5-based Retriever (GTR)

- **Data efficiency**: Only needs **10%** of MS Marco *supervised data* to achieve the best **out-of-domain** performance!

*GTR **w/o** Pre-training — *GTR w/ Pre-training + Fine-tuning

	GTR-FT		GTR		
Ratio of data	Large	XL	Large	XL	XXL
	NDCG@10 on MS Marco *in-domain				
10%	0.402	0.397	0.428	0.426	-
100%	0.415	0.418	0.430	0.439	0.430
	Zero-shot average NDCG@10 w/o MS Marco *out-of-domain				
10%	**0.413**	0.418	**0.452**	**0.462**	**0.465**
100%	0.412	**0.433**	0.445	0.453	0.458

Ni et al., 2021. Large Dual Encoders Are Generalizable Retrievers

Challenges in Modern RAG

- A significant amount of noise information even fake news in the content available on the Internet.
- Currently, there is a lack of comprehensive understanding of how each model can improve performance through information retrieval.

Type of Noises

- Relevant (semantically similar) but not contain the answer
- Counterfactual information
- Irrelevant information
- …

Capabilities that LLMs Should Have in RAG

Noise Robustness

- LLMs must be able to extract the necessary information from documents despite there are noisy documents.

Question

Who was awarded the **2022** Nobel prize in literature?

External documents contain noises

The Nobel Prize in Literature for **2022** is awarded to the French author **Annie Ernaux**, “for the courage and clinical acuity …

The Nobel Prize in Literature for **2021** is awarded to the novelist **Abdulrazak Gurnah**, born in Zanzibar and active in …

Retrieval Augmented Generation

Annie Ernaux

Capabilities that LLMs Should Have in RAG

Negative Rejection

- In real-world situations, the search engine often fails to retrieve documents containing the answers.
- It is important for the model to have the capability to reject recognition and avoid generating misleading content.

Question

Who was awarded the **2022** Nobel prize in literature?

External documents contain noises

The Nobel Prize in Literature for **2021** is awarded to the novelist **Abdulrazak Gurnah**, born in Zanzibar and active in …

The **202**0 Nobel Laureate in Literature, poet **Louise Glück**, has written both poetry and essays about poetry. Since her…

Retrieval Augmented Generation

I can not answer the question because of the insufficient information in documents.

Capabilities that LLMs Should Have in RAG

Information Integration

- In many cases, the answer to a question may be contained in multiple documents.
- To provide better answers to complex questions, it is necessary for LLMs to have the ability to integrate information.

Question

When were the **ChatGPT app for iOS** and **ChatGPT api** launched?

External documents contain noises

On **May 18th**, 2023, OpenAI introduced its own **ChatGPT app for iOS**…

That changed **on March 1**, when OpenAI announced **the release of API access to ChatGPT** and Whisper,…

Retrieval Augmented Generation

May 18 and March 1.

Capabilities that LLMs Should Have in RAG

Counterfactual Robustness

- In the real world, there is an abundance of false information on the internet.
- LLMs should identify risks of known factual errors in the retrieved documents when the LLMs are given warnings about potential risks in the retrieved information through instruction.

Question

Which city hosted the Olympic games in **2004**?

External documents contain noises

The 2004 Olympic Games returned home to **New York**, birthplace of the …

After leading all voting rounds, **New York** easily defeated Rome in the fifth and final vote …

Retrieval Augmented Generation

There are factual errors in the provided documents. **The answer should be Athens.**