Transformer-based models and Applications for Information Retrieval

NTUST - Information Retrieval (2020.12.18)

臺科大
TAIWAN TECH

郭家廷 **(Chia-Chih Kuo)**

Natural Language Processing Laboratory

National Taiwan University of Science and Technology

Transformer

BERT

XLNet

RoBERTa

ALBERT

Tokenization

Homework 6

Attention mechanism

✓ Attention mechanism

- Encodes long-distance dependency
- Captures contextual relationship
- Widely used in a variety of neural networks

Attention mechanism

✓ Attention mechanism

…is just a fancy word for **weighted average**

α_i	0.04	0.14	0.06	0.73	0.01	0.02		
	They	eat	two	apples	this	morning	<query>	red fruit
w_i	(300-dim embedding for each word)							

- **Conceptually,** $\text{Attention} = \alpha_i w_i$

Attention mechanism

✓ Scaled dot-product attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

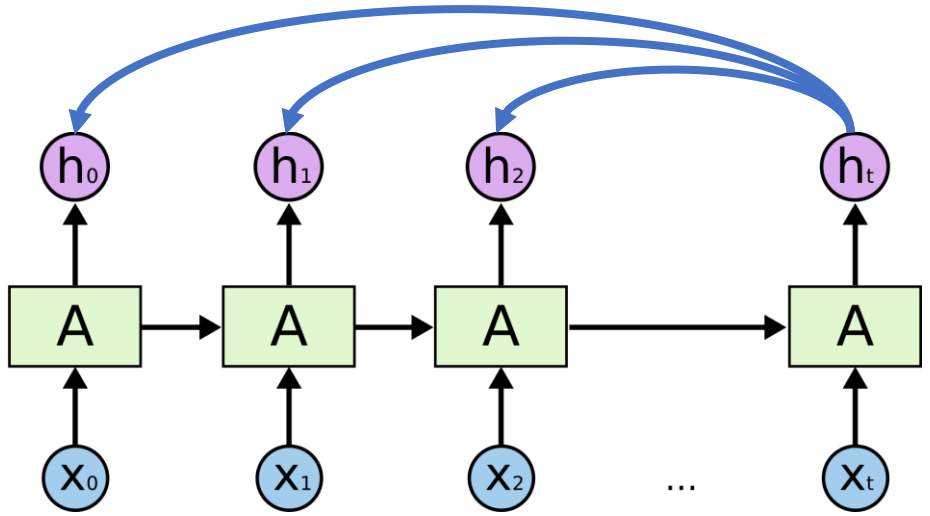
Feed query/document through individual FFNs to get Q, K, V:

- $W_{\text{query}}, W_{\text{doc}}$: a sequence of word embeddings of query/document
- d_k : dimension of word embeddings
- $Q = \text{FFN}_Q(W_{\text{query}})$; $K = \text{FFN}_K(W_{\text{doc}})$; $V = \text{FFN}_V(W_{\text{doc}})$

Attention mechanism

✓ RNN with attention

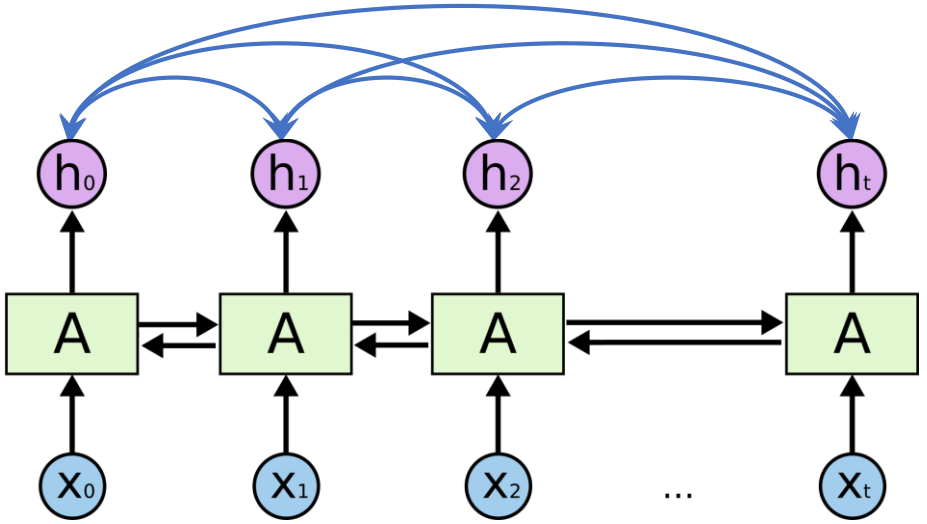
- Usually take the last hidden output as query
- Performance degrades if the distance is very long



Attention mechanism

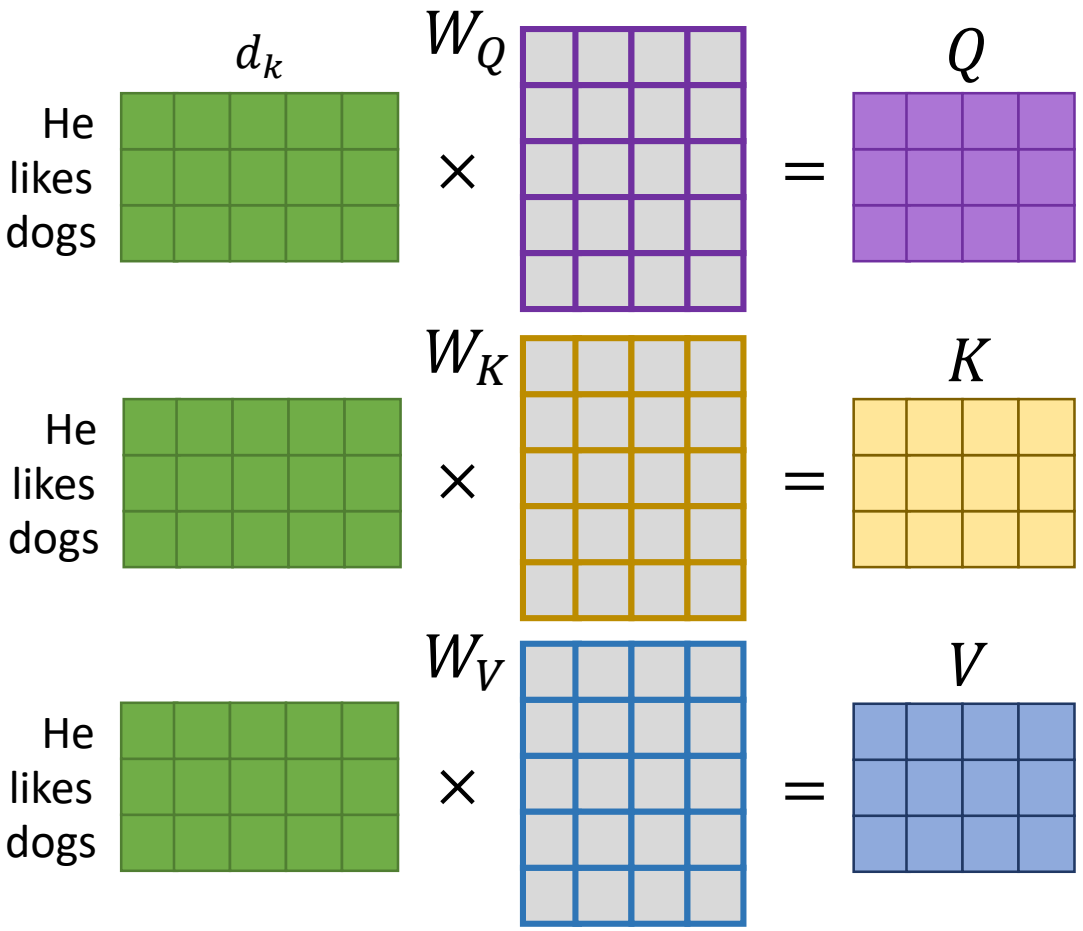
✓ Bidirectional RNN with self-attention

- Attention of all possible pairs of any hidden outputs
- Alleviate but still suffer for long-distance problem

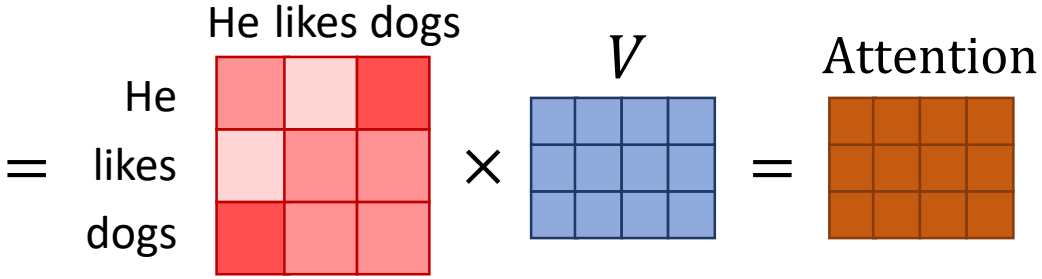


Transformer

✓ Self-attention



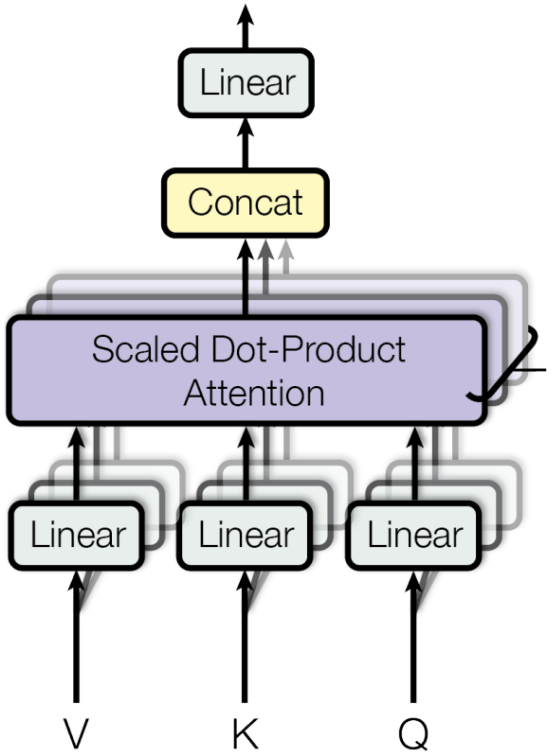
$$\text{softmax}\left(\frac{Q \times K^{\top}}{\sqrt{d_k}}\right) V$$



Transformer

✓ Multi-head attention

- Combine multiple attention layers **in parallel**
- Increase feature resolution
- Similar effect of ensemble



Transformer

✓ Transformer

- The first model entirely relies on self-attention
- No recurrent nor convolution operations

Layer Type	Complexity per Layer	Sequential Operations
Self-Attention	$O(\text{length}^2 \cdot \text{dim})$	$O(1)$
Recurrent	$O(\text{length} \cdot \text{dim}^2)$	$O(\text{length})$
Convolutional	$O(\text{kernel} \cdot \text{length} \cdot \text{dim}^2)$	$O(1)$

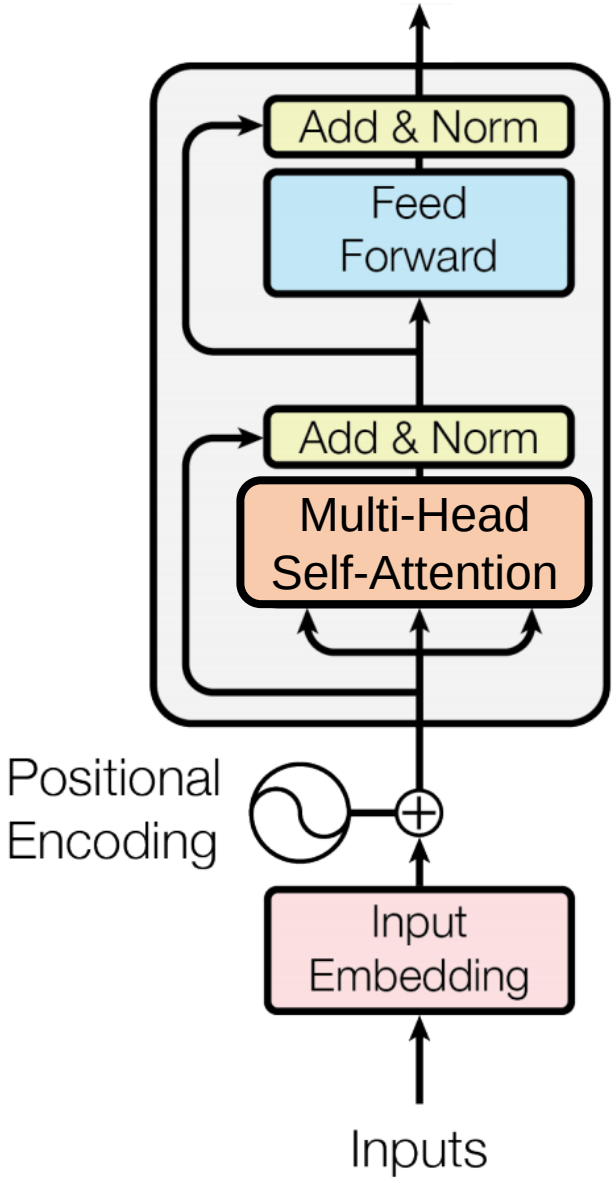
length: input length; dim: hidden size; kernel: kernel size

Transformer

✓ Transformer encoder

- ➢ A multi-head self-attention layer
- ➢ Residual connections
- ➢ Layer normalization

$$x = \text{Emb}(\text{inputs}) + \text{Emb}(\text{position})$$
$$z = \text{LayerNorm}\big(x + \text{MHSA}(x)\big)$$
$$y = \text{LayerNorm}(z + \text{FFN}(z))$$



Transformer

BERT

XLNet

RoBERTa

ALBERT

Tokenization

Homework 6

BERT

✓ **BERT** -- Jacob Devlin et al., October 2018.

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Pretrain representations from unlabeled text
- Can be easily finetuned for a wide range of tasks
- Obtained new state-of-the-art results on 11 NLP tasks

BERT

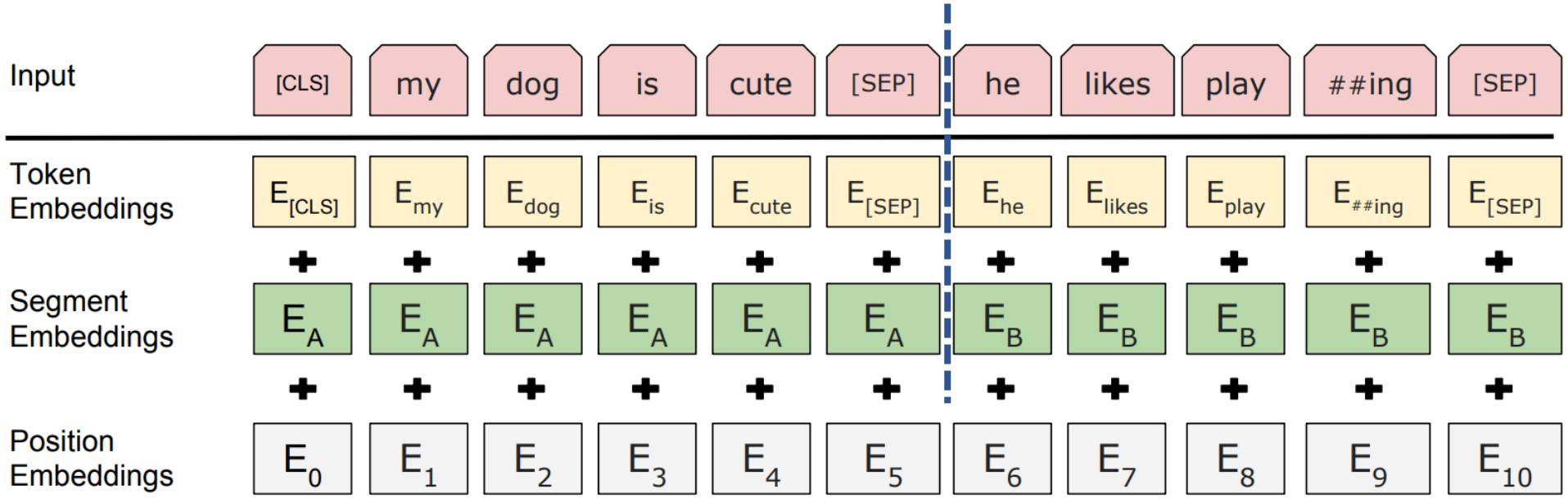
✓ **BERT** -- Jacob Devlin et al., October 2018.

- Consists of stacked transformer layers and 3 embedding layers
- Officially provide multiple size of pretrained BERT
- BERT-base: 12 layers, 12 heads, 768-hidden (110M params)
- BERT-large: 24 layers, 16 heads, 1024 hidden (340M params)

BERT

✓ **BERT** -- Jacob Devlin et al., October 2018.

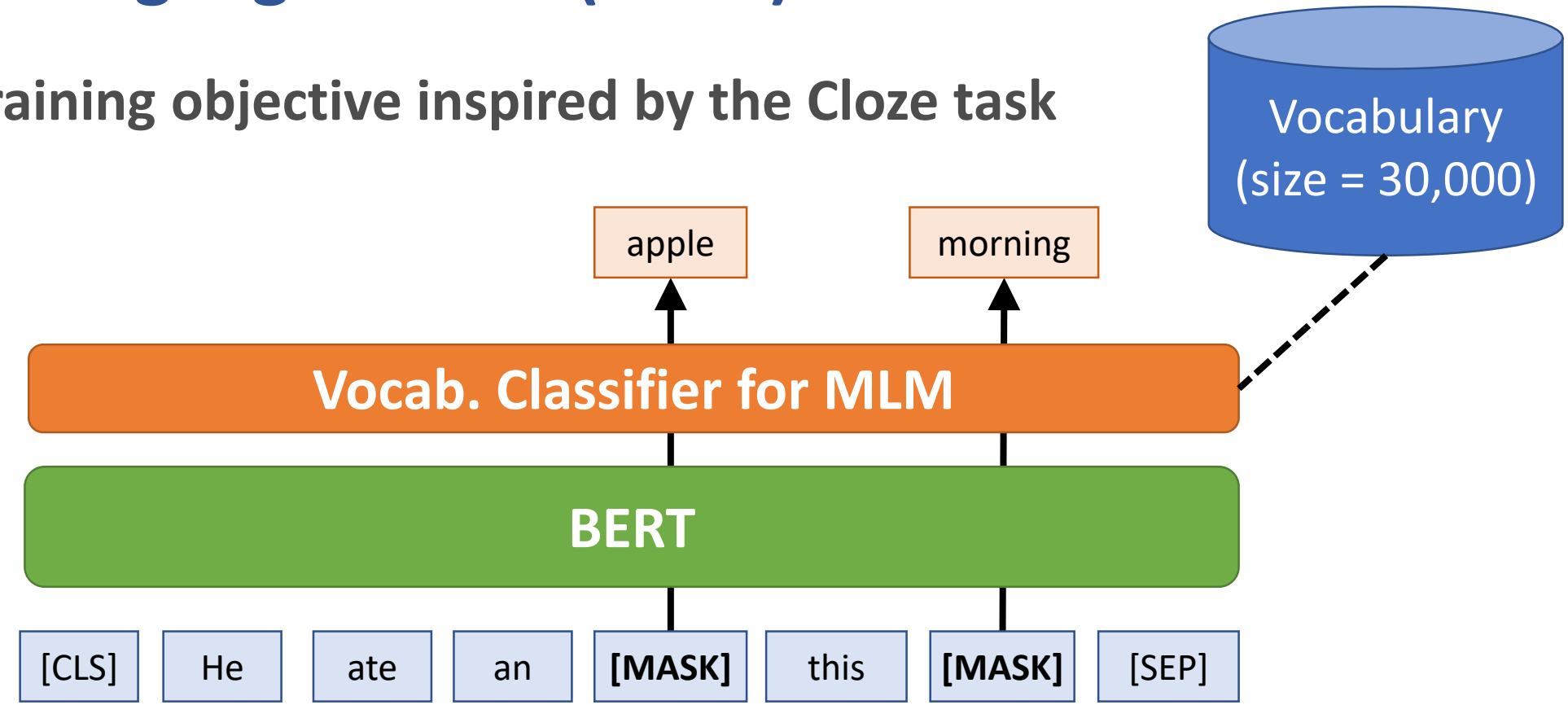
- Use **WordPiece** tokens (a subword tokenization method)
- Sum up token/segment/position embeddings as input



BERT

✓ Masked Language Model (MLM)

- A pretraining objective inspired by the Cloze task



BERT

✓ Masked Language Model (MLM)

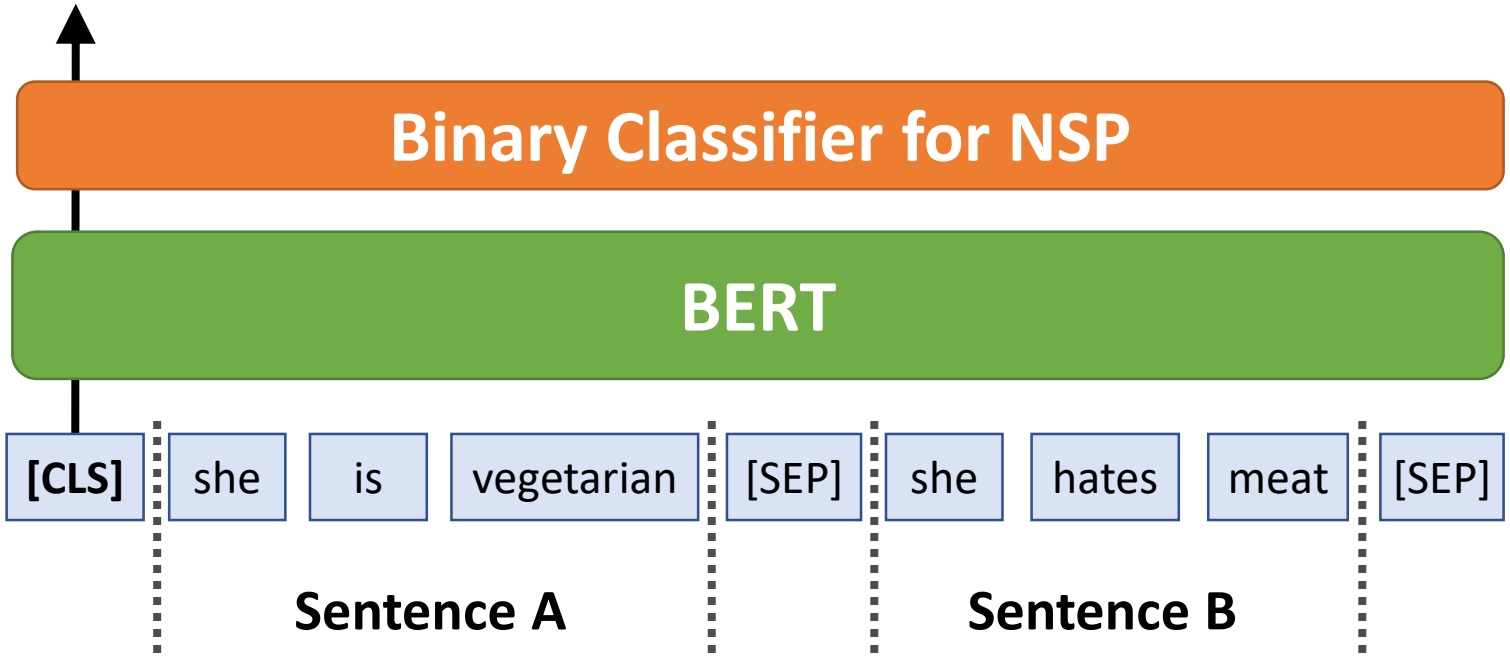
- **A pretraining objective inspired by the Cloze task**
- **Randomly mask 15% tokens in each sequence**
  - 80% are replaced with [MASK] token
  - 10% are replaced with other random tokens
  - 10% are keep unchanged
  - Masking and replacement is performed once in the beginning

BERT

✓ Next Sentence Prediction (NSP)

- **Pretraining for understanding of sentence-level relationship**

label = **1** if **Sentence B** is the actual next sentence of **Sentence A**, else **0**



BERT

✓ Next Sentence Prediction (NSP)

- **Pretraining for understanding of sentence-level relationship**
- **Extract contiguous sequences from document-level corpus**
  - 50% of Sentence B is the actual next sentence of sentence A
  - 50% of Sentence B is a random sentence from the corpus

BERT

✓ Pretraining of BERT

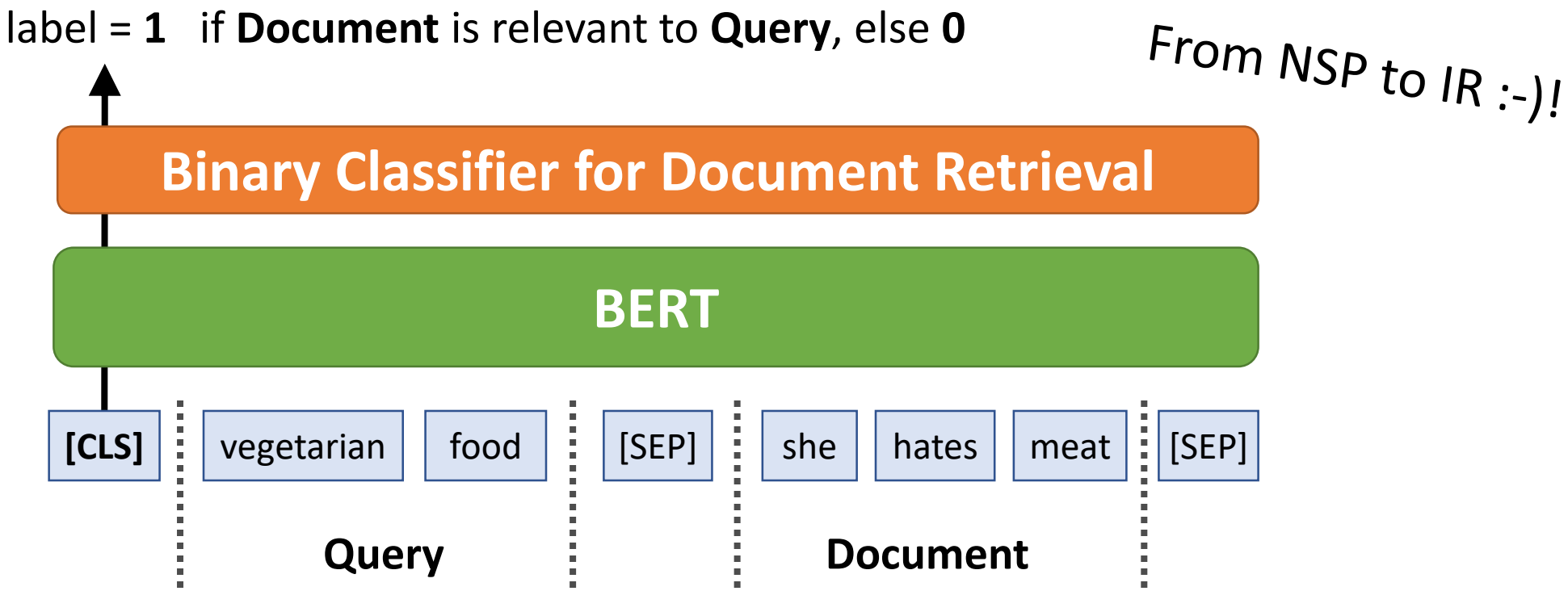
- Jointly pretrain MLM and NSP objectives
- Officially pretrain on BooksCorpus (800M words) & Wikipedia (2.5B words)
- Classifiers in MLM and NSP are only used for pretraining

  (i.e. you should stack a new classifier on BERT for downstream tasks)

BERT

✓ Finetune BERT on downstream tasks

- **Highly compatible due to its architecture and pretraining**



Transformer

BERT

XLNet

RoBERTa

ALBERT

Tokenization

Homework 6

AR & AE

✓ Autoregressive (AR)

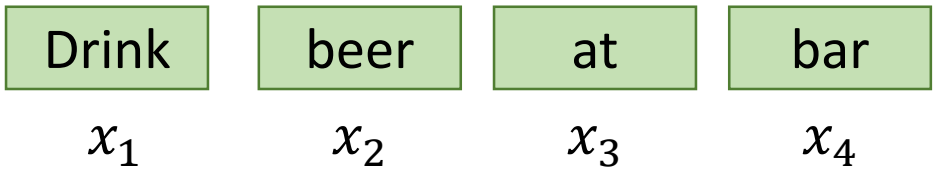
- **Autoregressive language modeling:** (unidirectional)

  Predict x_3 with $\{x_1, x_2\}$; predict x_4 with $\{x_1, x_2, x_3\}$

✓ Autoencoding (AE)

- **Denoising autoencoding (DAE):** data reconstruction (bidirectional)

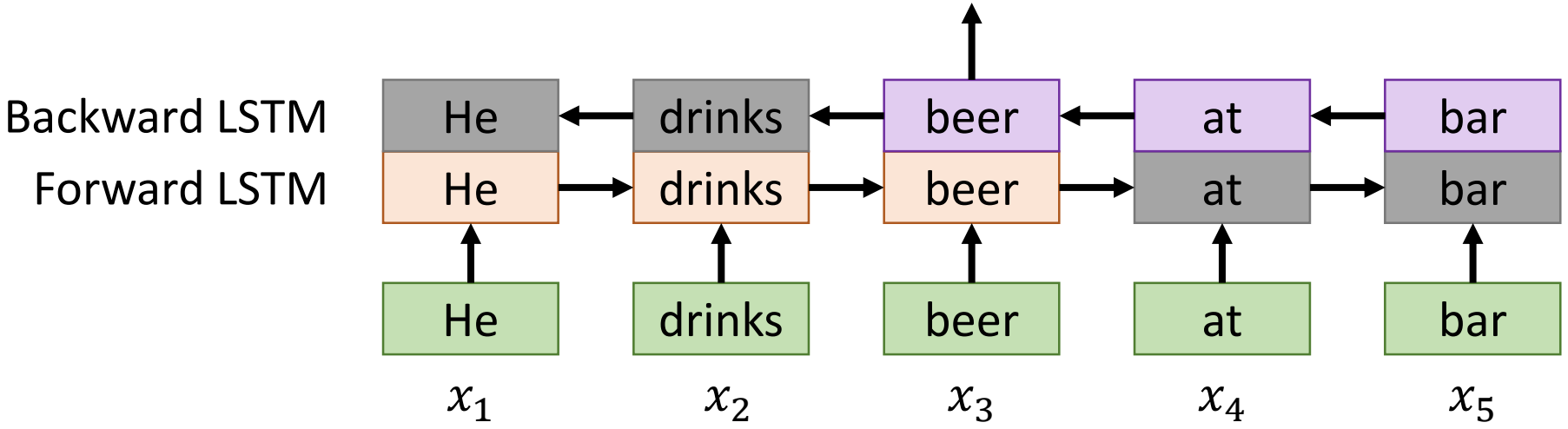
  Predict x_3 with $\{x_1, x_2, x_4\}$; predict x_4 with $\{x_1, x_2, x_3\}$



AR & AE

✓ **ELMo** -- Matthew E. Peters et al., March 2018.

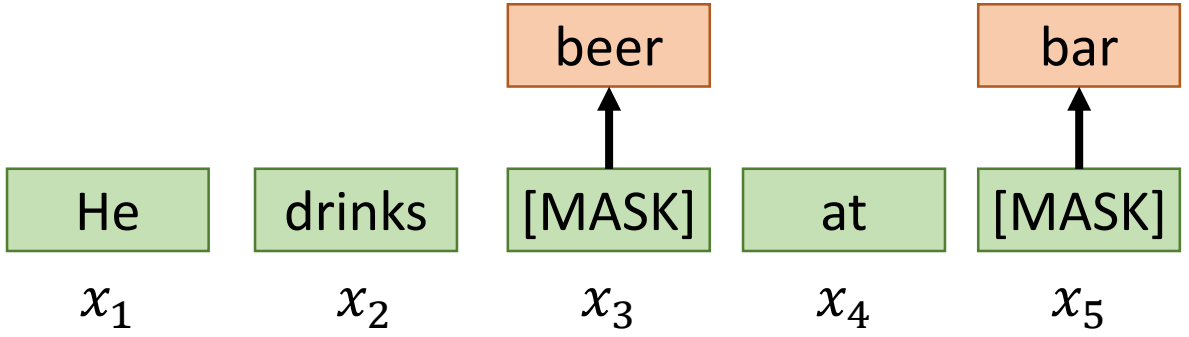
- **AR language modeling with bidirectional LSTM**
- **Base layer only encodes unidirectional information**



AR & AE

✓ MLM is a DAE-based pretraining

- Has capability of modeling bidirectional contexts
- BERT outperforms AR-based models like ELMo, GPT
- So… what's wrong with MLM?



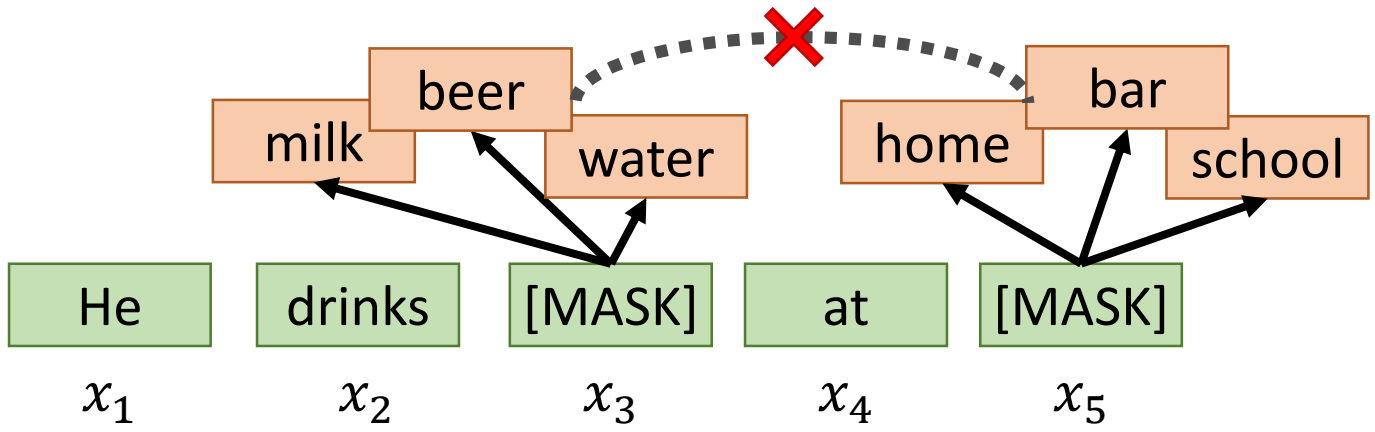
AR & AE

✓ MLM is a DAE-based pretraining

- **Predict $\{x_3, x_5\}$ with $\{x_1, x_2, x_4\}$:**

  → unable to model the dependency between x_2 and x_4

- **[MASK] token is never used in downstream tasks → input noise**



XLNet

✓ **XLNet** -- Zhilin Yang et al., June 2019.

- A novel AR pretraining method which learns bidirectional context
- Overcomes the limitations of MLM thanks to AR formulation
- Not using [MASK] token for pretraining
- Outperforms BERT on 20 NLP tasks

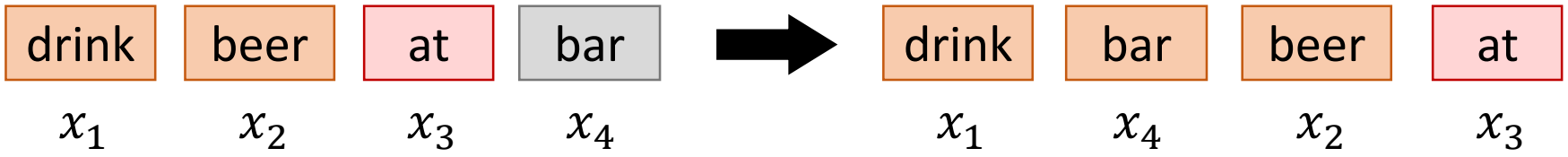
In brief, XLNet has both the advantages of AR and AE

XLNet

✓ Permutation Language Modeling (PLM)

- **AR pretraining with all possible permutations of a sequence**

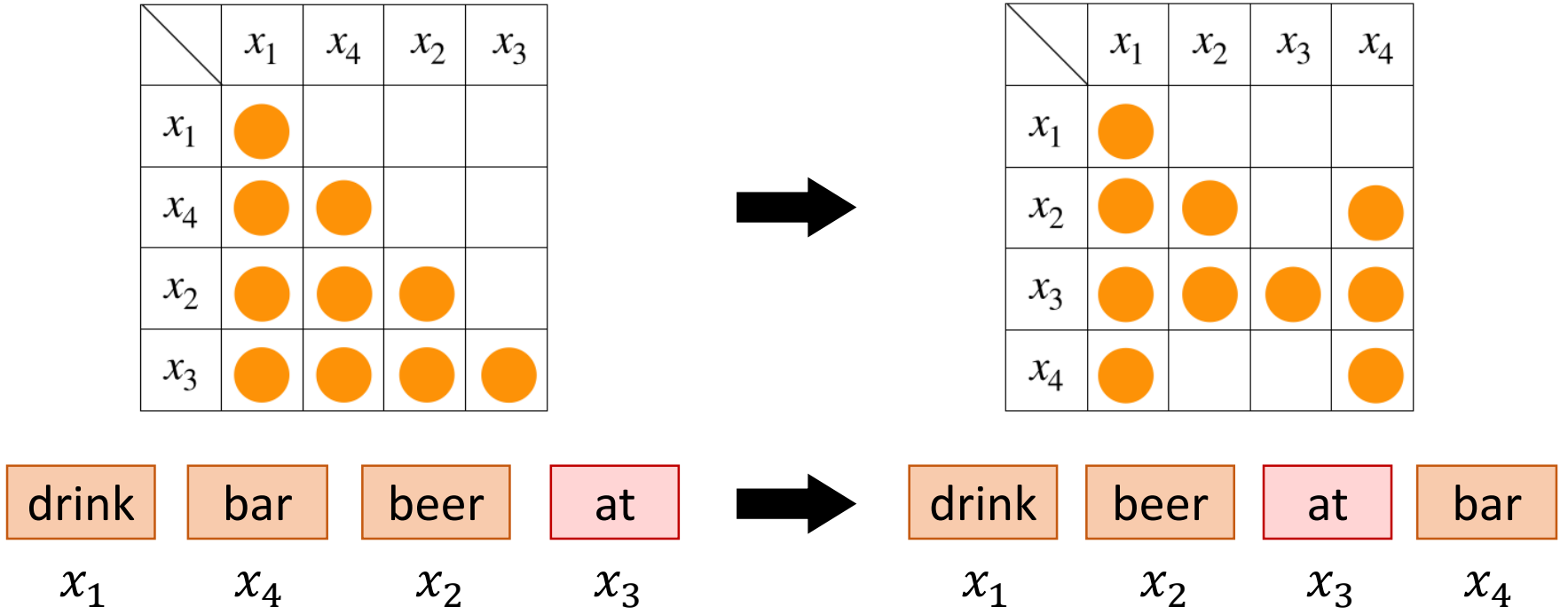
1. Permutate {**1, 2, 3**, 4} → predict x_3 with $\{x_1, x_2\}$ (regular AR)
2. Permutate {**2, 4, 3**, 1} → predict x_3 with $\{x_2, x_4\}$
3. Permutate {**1, 4, 2, 3**} → predict x_3 with $\{x_1, x_4, x_2\}$
4. …etc



XLNet

✓ Permutation Language Modeling (PLM)

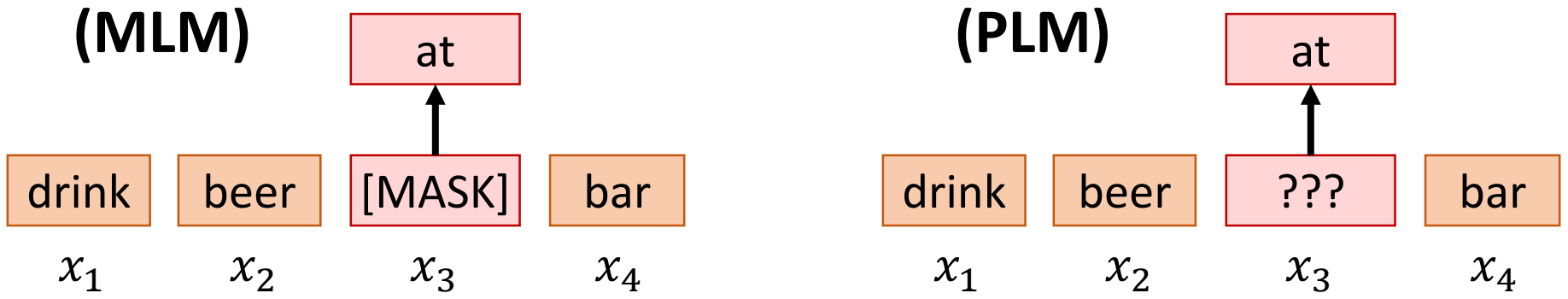
- **Autoregressive self-attention mask**



XLNet

✓ Permutation Language Modeling (PLM)

- AR now learns bidirectional contexts with input permutation
- The token classifier shall not see the target token!
- So… how to mask the target token without using [MASK] token?

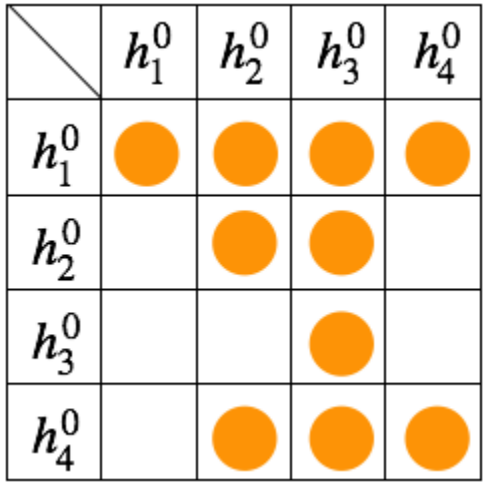


XLNet

✓ Two-Stream Self-Attention

- Assume we have an AR self-attention mask,

  and we want to predict x_1 with $\{x_2, x_3, x_4\}$

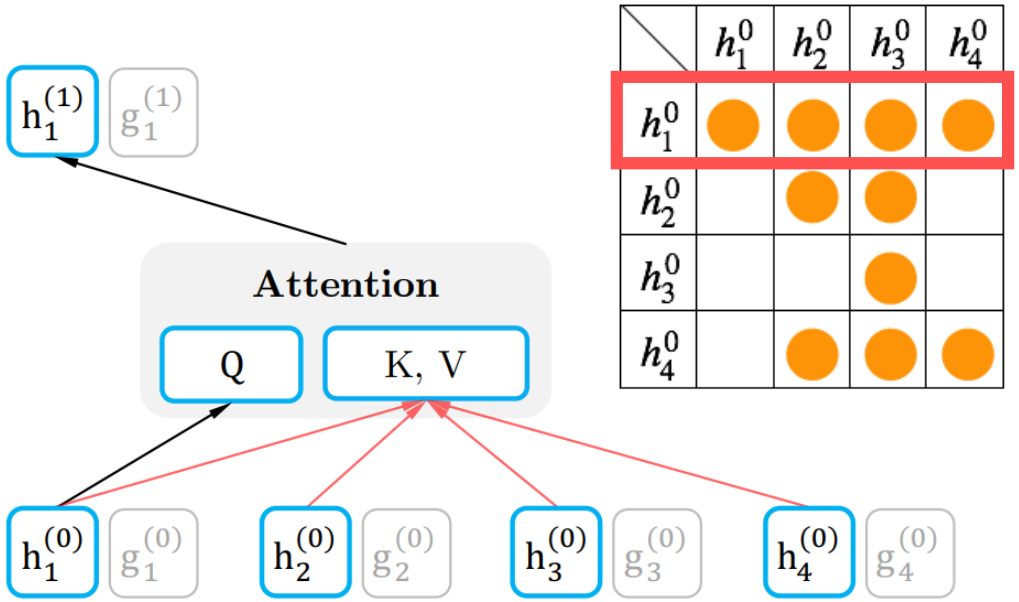


XLNet

✓ Two-Stream Self-Attention

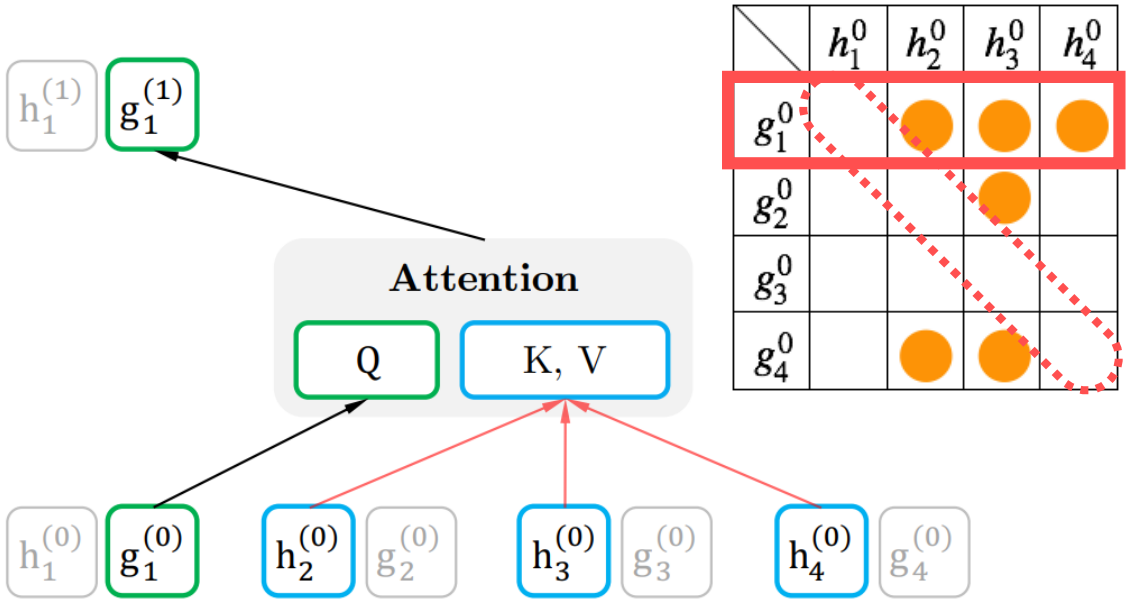
Content Stream:

can see self

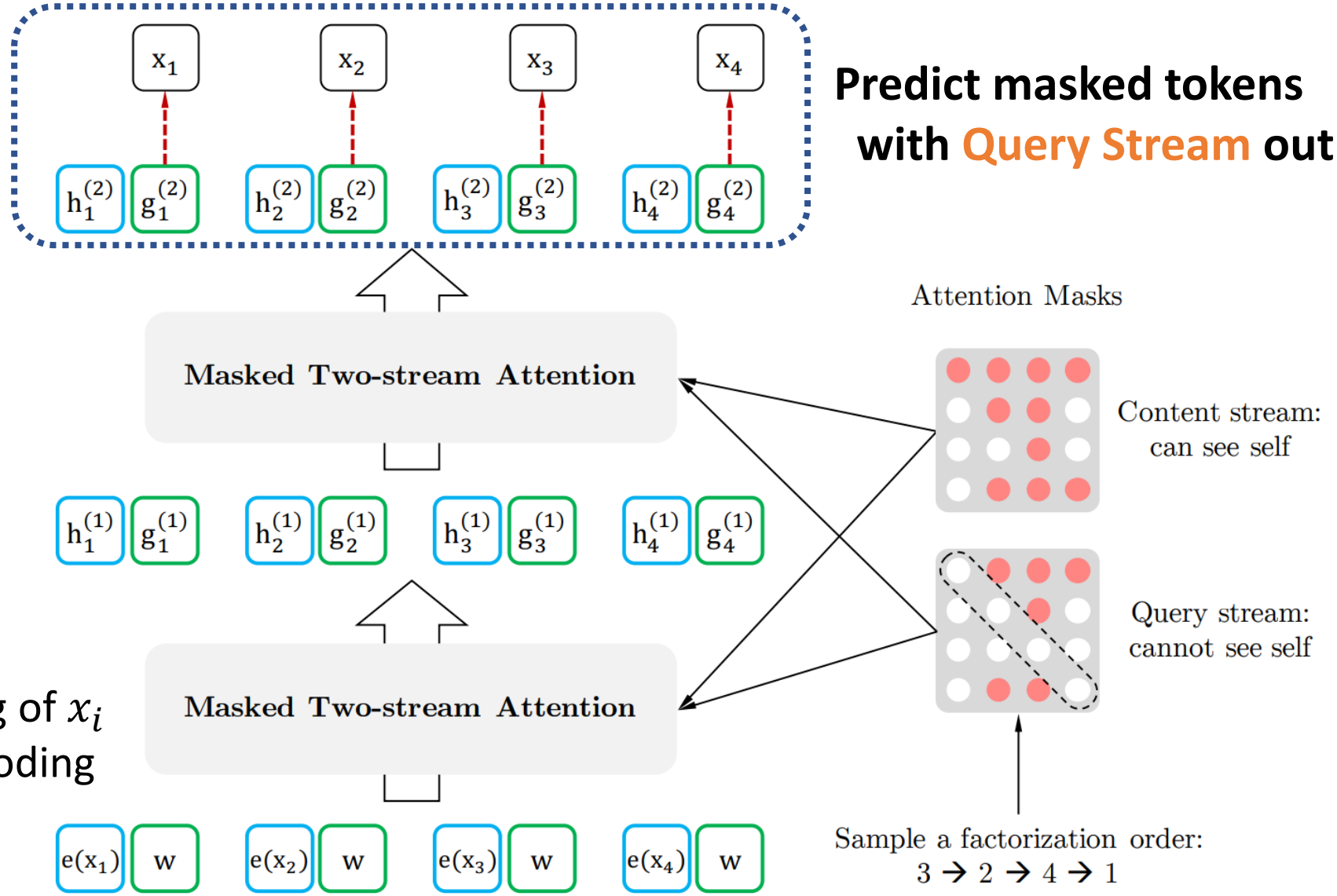


Query Stream: (for PLM only)

cannot see self



XLNet



Predict masked tokens with Query Stream outputs

$e(x_i)$: embedding of x_i
w : positional encoding

XLNet

✓ Pretraining of XLNet

- Pretrain PLM objective only (NSP is found unhelpful for XLNet)
- XLNet uses SentencePiece tokenization

10x MORE DATA ;-) !

- Officially pretrain on BooksCorpus (800M words), Wikipedia (2.5B words), Giga5 (16GB text), ClueWeb 2012-B, and Common Crawl
- Query Stream is only used in PLM;

  Content Stream is used in PLM and downstream tasks

Transformer

BERT

XLNet

RoBERTa

ALBERT

Tokenization

Homework 6

RoBERTa

✓ **RoBERTa** -- Yinhan Liu et al., July 2019.

- Robustly optimized BERT approach
- Original BERT (by Google) is significantly undertrained
- Propose an improved recipe for training BERT models
- Match or outperform all post-BERT methods (e.g. XLNet)

RoBERTa

	BERT	RoBERTa
Batch size	256	**8K**
Training steps	1M	500K
Corpus size	13GB	**160GB**
MLM masking	Static (fixed once prepared)	Dynamic (generate on the fly)
Objective	MLM + **NSP**	MLM
Tokenization	WordPiece (vocab. size = 30k)	Byte-level **BPE** (vocab. size = 50k)
⋮		

RoBERTa

✓ RoBERTa

- **Significantly outperforms BERT on GLUE benchmark**
- **Outperforms XLNet on every single task in GLUE**

* General Language Understanding Evaluation (GLUE) benchmark*

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-

Transformer

BERT

XLNet

RoBERTa

ALBERT

Tokenization

Homework 6

ALBERT

✓ **ALBERT** -- Yinhan Liu et al., September 2019.

- A Lite BERT
- Difficult to experiment with large models due to memory constraints
- Propose several parameter-reduction techniques
- Propose a new pretraining method to replace NSP

ALBERT

✓ Factorized embedding parameterization

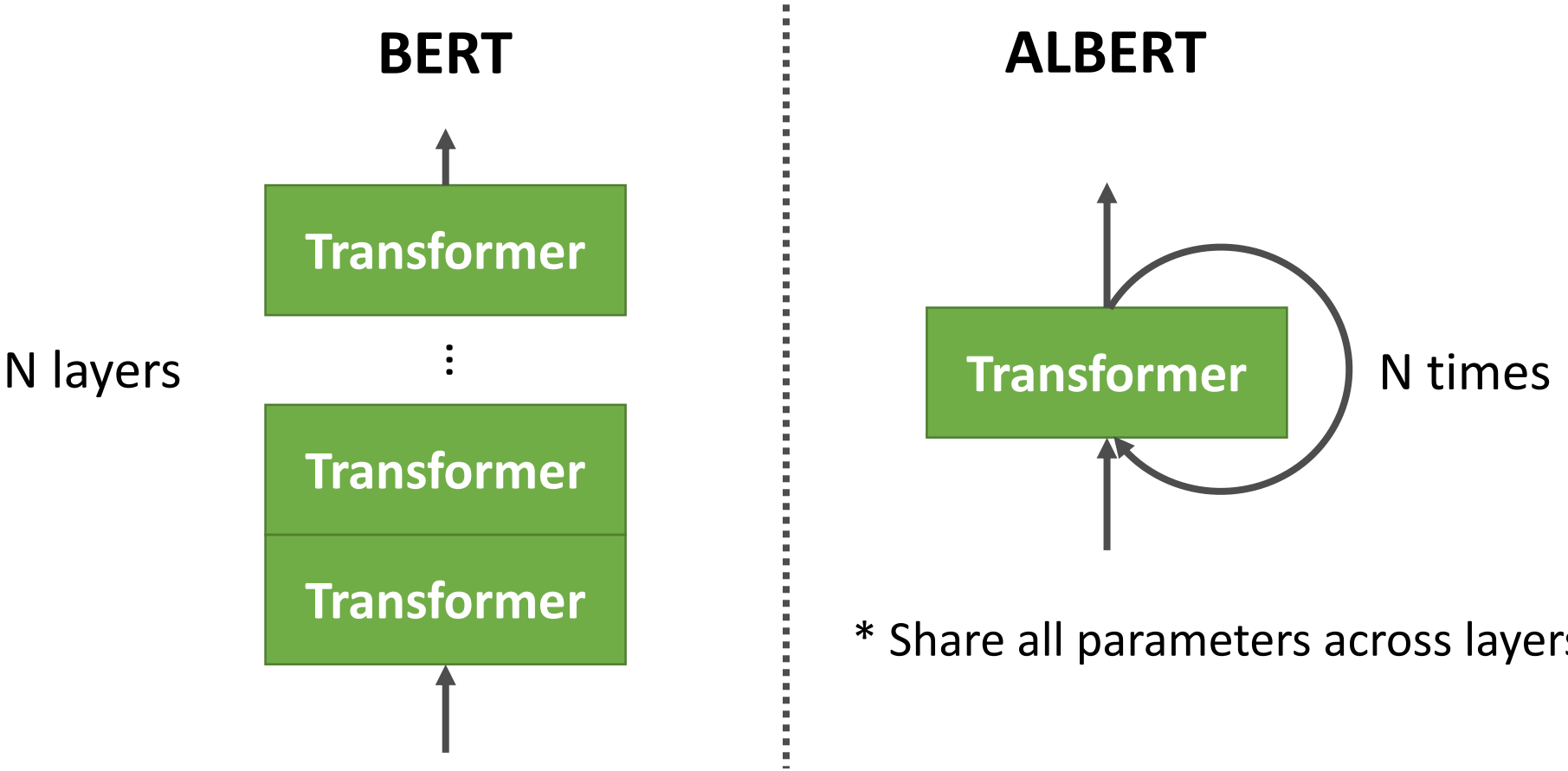
- Word embeddings are meant to be context-independent
- Hidden outputs are meant to be context-dependent
- It is efficient and reasonable to have a smaller word dimension

Both BERT and ALBERT have vocabulary size = 30,000:

- BERT = Embedding(30000, 768) (23M params)
- ALBERT = Embedding(30000, 128) → FFN(128, 768) (4M params)

ALBERT

✓ Cross-layer parameter sharing



ALBERT

✓ Why is Next Sentence Prediction (NSP) unhelpful?

- **NSP conflates topic prediction and coherence prediction**
- **It is much easier to learn topic prediction (high word overlapping)**
- **Models tend to learn the easier topic-prediction signal**

	Sentence A	Sentence B
(Positive)	I play with my dog.	My dog makes me exhausted.
(Negative)	I play with my dog.	She drank a glass of water.

ALBERT

✓ Sentence Order Prediction (SOP)

- **Swap the order of positive samples from NSP as negative samples**
- **Force models to learn coherence prediction** **(topics are unchanged)**

	Sentence A	**Sentence B**
(Positive)	I play with my dog.	My dog makes me exhausted.
(Negative)	My dog makes me exhausted.	I play with my dog.

ALBERT

✓ Pretraining of ALBERT

- **Jointly pretrain MLM and SOP objectives**
- **ALBERT uses SentencePiece tokenization**
- **Officially pretrain on BooksCorpus (800M words) & Wikipedia (2.5B words)**

* to fairly compare with BERT *

ALBERT

✓ Parameter-efficiency of ALBERT

- ALBERT can have **significantly fewer parameters**

  and faster speed without seriously hurting performance

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

Transformer

BERT

XLNet

RoBERTa

ALBERT

Tokenization

Homework 6

Tokenization

✓ Vocabulary issues

- Common vocab. size= 50k~200k for TFIDF/RNN-based models
- Millions of unique words in a big corpus like Wikipedia!
- Out-of-vocabulary (OOV) words ALWAYS exists! (e.g. internet slang)
- Subword: seek a trade-off between semantics and memory

Tokenization

✓ **Byte Pair Encoding** -- Philip Gage, February 1994.

- A data compression algorithm
- Algorithm flow:

1. Vocabulary initialized with all unique characters
2. Greedily merge bigram with highest frequency
3. Stop merging if conditions are met (e.g vocab. size = 30k)

Tokenization

✓ Byte Pair Encoding -- Philip Gage, February 1994.

1. **{** “l o w”: **5** , “l o w e r”: **2** , “n e w e s t”: 6 , “w i d e s t”: **3 }**

   → **Bigram freq. =** {**es: 9** , st: 9 , **we: 8** , lo: 7 , ow: 7 , ne: 6 , … }

2. **{** “l o w”: **5** , “l o w e r”: **2** , “n e w **es** t”: **6** , “w i d **es** t”: **3 }**

   → **Bigram freq. =** {**est: 9** , lo: 7 , ow: 7, ne: 6, … , **we: 2** , … }

3. **{** “l o w”: **5** , “l o w e r”: **2** , “n e w **est**”: **6** , “w i d **est**”: **3 }**

   → **Bigram freq. =** {lo: 7 , ow: 7, ne: 6, … , we: 2 , … }

……

Tokenization

✓ **WordPiece** -- Mike Schuster, March 2012.

- A variation of BPE with adaptation to language modeling
- Merge bigram which increases likelihood the most on the corpus

1. { “l o w”: **5** , “l o w e r”: **2** , “n e w e s t”: 6 , “w i d e s t”: **3** }

   → Δ **Likelihood =** {es: 0.14 , st: 0.14 , we: 0.12 , lo: 0.11 … }

...

Tokenization

✓ **Unigram Language Modeling (Uni. LM)** -- Taku Kudo, April 2018.

Algorithm flow:

1. **Vocabulary initialized with all possible substrings**
2. **Prune 20% tokens which decrease likelihood the most**
3. **Stop pruning if conditions are met**

1. lower, … → {l, o, w, e, r, lo, ow, we, er, …, low, owe, …, lowe, ower, lower…}

   → Δ **Likelihood =** { ower: −2.7 , lo: −1.3 , … , we: 0.12 , low: 0.11 … }

…

Tokenization

✓ **BPE** with vocab. = {dis, car, de, ed, d}

discarded → dis | car | de | d * Greedy substitution *

✓ **Uni. LM** * Viterbi decoding for **maximum likelihood** *

		0.0005	
discarded	→	**discarded**	**p = 0.0005**
		0.006 0.3	
discarded	→	**discard | ed**	**p = 0.0018**
		0.1 0.1 0.4 0.3	
discarded	→	**dis | car | d | ed**	**p = 0.0012**
		0.1 0.1 0.2 0.4	
discarded	→	**dis | car | de | d**	**p = 0.0008**
		0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4	
discarded	→	**d | i | s | c | a | r | d | e | d**	**p ≈ 0.00026**

Tokenization

✓ **SentencePiece** -- Taku Kudo et al., August 2018.

- A open-sourced library made by Google
- Take whole sentences to train BPE/Uni. LM

e.g. SentencePiece with BPE

1. { “t h i s _ i s _ l o w e r”: **3** , “w h a t _ i s _ t h e _ n e w e s t”: **2** }

   → **Bigram freq. =** { is: 8 , th: 5 , we: 5 , … }

…

Tokenization

✓ BERT	WordPiece
✓ RoBERTa, GPT-2	BPE over raw bytes
✓ T5	BPE over Unicode characters
✓ XLNet, ALBERT	SentencePiece

- Both papers don’t clarify whether they use BPE or Uni. LM

Transformer

BERT

XLNet

RoBERTa

ALBERT

Tokenization

Homework 6

Homework 6

✓ Homework 6

- Goal: Rescore BM25 retrieval with Transformer-based models
- Performance are measured with MAP@1000
- It is NOT allowed to use any approach

  that doesn't involve a Transformer-based model!

e.g. X Simply finetune a strong BM25 system

e.g. O Strong BM25 + weak BERT (allowed but not encouraged)

e.g. Δ Strong BM25 = 45.08 → + BERT = 43.78 (harmful BERT rescoring)

Homework 6

✓ Homework 6

Given data:

- **100,000 documents**
- **120 training queries & 80 testing queries (50% for private)**
- **Positive (i.e. relevant) document IDs for each training query**
- **Top-1000 BM25 document IDs and scores for all queries**

Homework 6

✓ Homework 6

Data Explorer

336.33 MB

- documents.csv
- sample_submission.csv
- test_queries.csv
- train_queries.csv

documents.csv (331.49 MB)

Detail | Compact | Column | 2 of 2 columns

About this file

This file contains all the documents for information retrieval.

doc_id Document ID	doc_text Document text
100000 unique values	**99073** unique values
FBIS3-100	Language: <F P=105> English </F> Article Type:BFN [Text] Nairobi, 28 Feb (KNA) -- The chai...
FBIS3-10005	Language: <F P=105> Spanish </F> Article Type:BFN [Text] Havana, 25 Feb (DPA) -- Today, Cu...
FBIS3-10007	Language: <F P=105> Spanish </F> Article Type:BFN <F P=106> [From the "Evening Information R...
FBIS3-10009	Language: <F P=105> Spanish </F> Article Type:BFN <F P=106> ["Stenographic version" of state...

Homework 6

✓ Homework 6

Data Explorer

336.33 MB

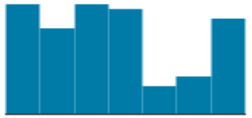
- documents.csv
- sample_submission.csv
- test_queries.csv
- train_queries.csv

train_queries.csv (2.93 MB)

Detail | Compact | Column

About this file

This file contains queries, positive (relevant) and BM25 top-1000 document IDs for training.

query_id Query ID	query_text Query text	pos_doc_ids Space-delimited document IDs which are relevant to the query	bm25_top1000 Space-delimited top-1000 document IDs ranked by BM25 (far left one is top-1)	bm25_top1000_s... Corresponding scores for "bm25_top1000" column
 302 650	120 unique values	120 unique values	120 unique values	120 unique values
302	Poliomyelitis and Post-Polio	FBIS3-20548 FBIS3-22539 FBIS3-22560 FBIS3-22589 FBIS3-26593 FBIS3-41672 FBIS3-41724 FBIS3-60403 FBIS...	FBIS4-67701 LA043090-0036 LA031489-0032 FBIS4-30637 FR940126-2-00106 FBIS3-60405 LA072890-0066 FBIS3...	32.84784386 31.61291462 23.97999093 22.78184683 19.16614366 18.75011496 18.37276912 18.03671536 18.0...
305	Most Dangerous Vehicles	FT922-1008 FT922-8544 FT944-136 FT944-15615 FT944-5300 FT944-9371 LA012590-0229 LA020490-0021 LA0205...	FBIS4-55132 FBIS3-56026 LA102790-0086 LA011490-0142 LA031689-0177 LA042690-0137 FBIS4-66625 LA021689...	16.54545283 16.40211005 16.06694476 15.48911938 15.24177722 14.76031153 14.52218082 14.35904320 14.3...

Homework 6

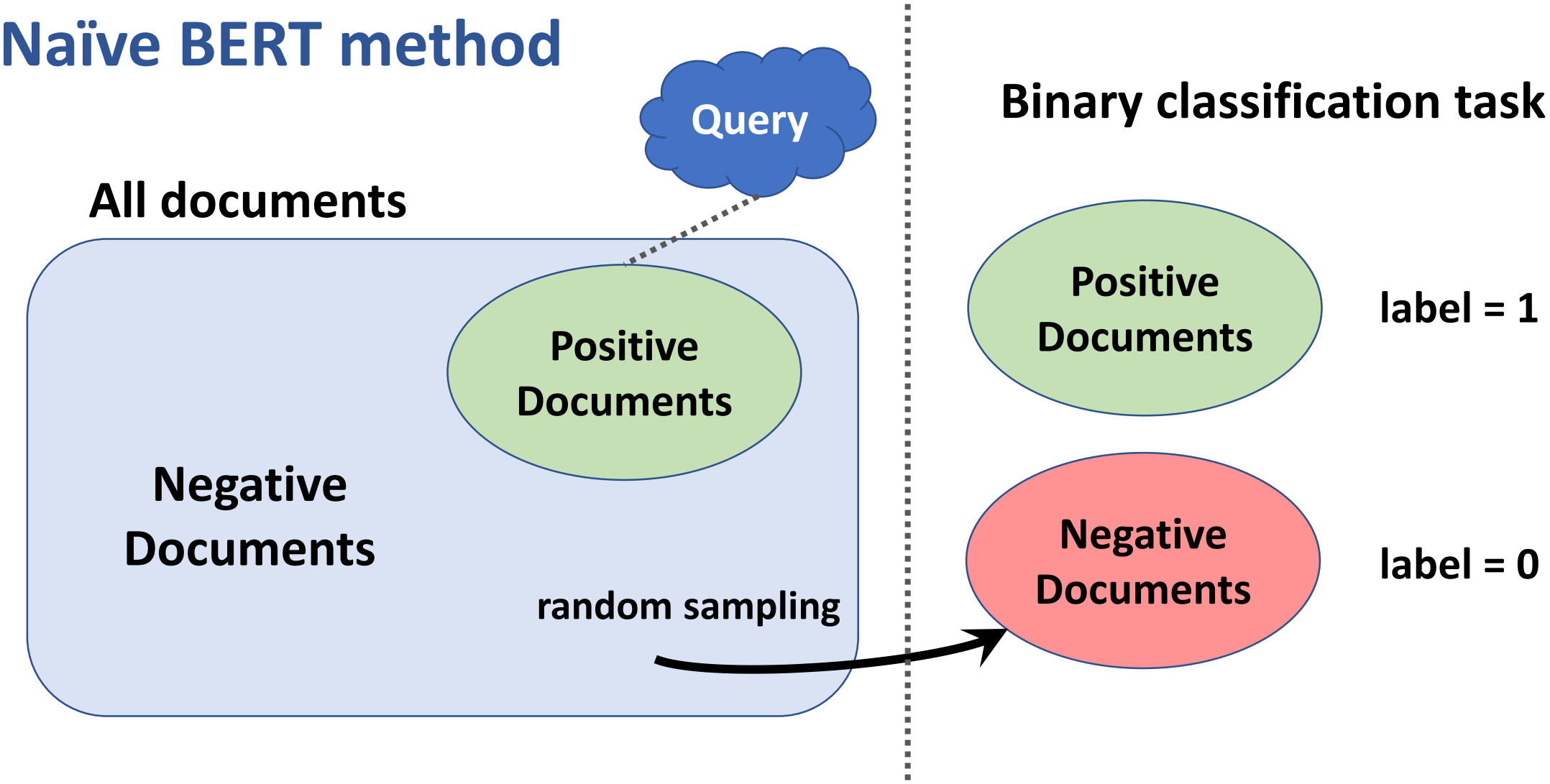
✓ Homework 6

Document IDs/scores are saved as space-delimited strings:

- **query_id** = 302
- **pos_doc_ids** = "FBIS3-20548 FBIS3-22539 FBIS3-22560 FBIS3-22589 FBIS3-26593 FBIS3-41672 FBIS3-41724 FBIS3-60403 ..."
- **bm25_top1000** = "FBIS4-67701 LA043090-0036 LA031489-0032 FBIS4-30637 FR940126-2-00106 FBIS3-60405 ... "
- **bm25_top1000_scores** = "32.84784386 31.61291462 23.97999093 22.78184683 19.16614366 18.75011496 ..."

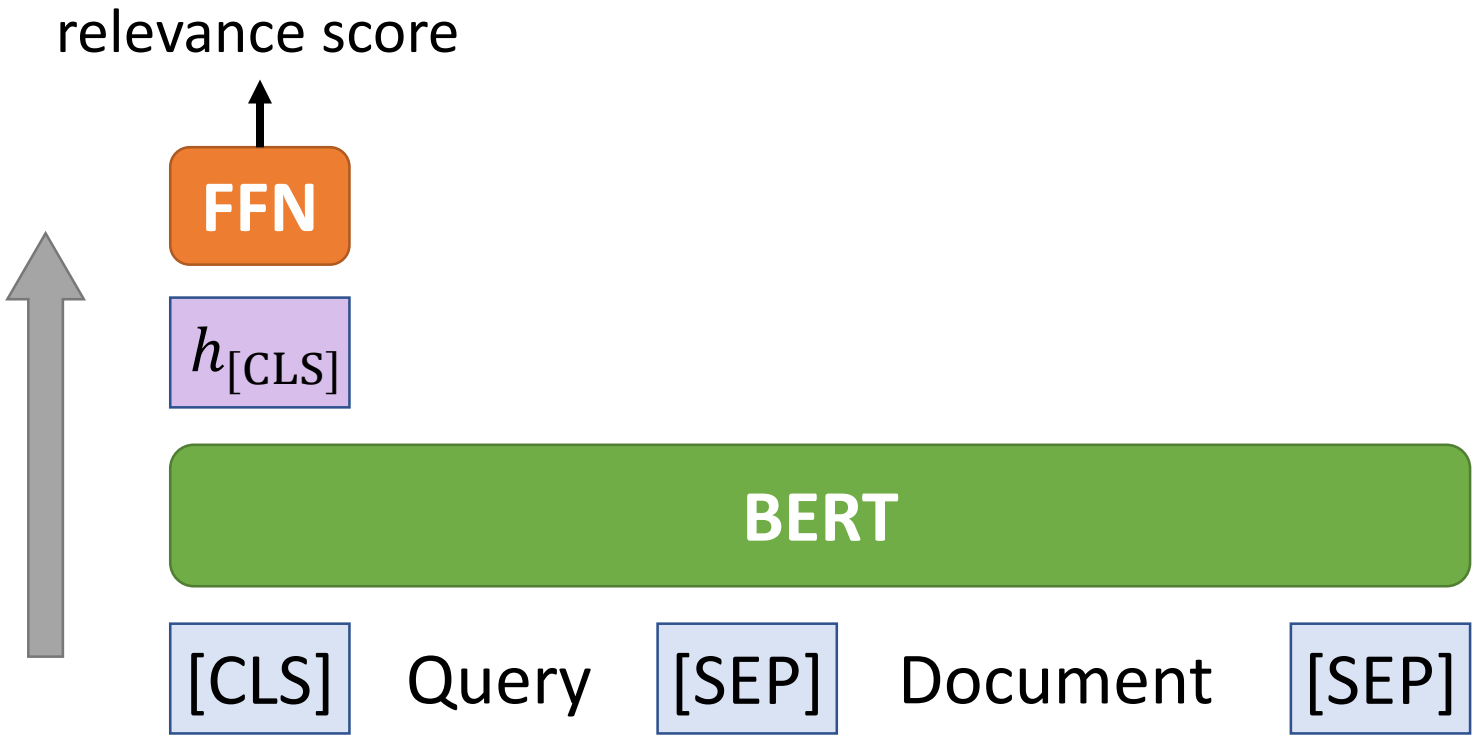
Homework 6 – 2019 Baseline

✓ Naïve BERT method



Homework 6 – 2019 Baseline

✓ Naïve BERT method



Homework 6 – 2019 Baseline

✓ Problems of naïve BERT method

- **Need to score all documents for a single query**

  Testing set: 80 queries * 100,000 docs = 8,000,000 computations

- **Lack of difficult samples**

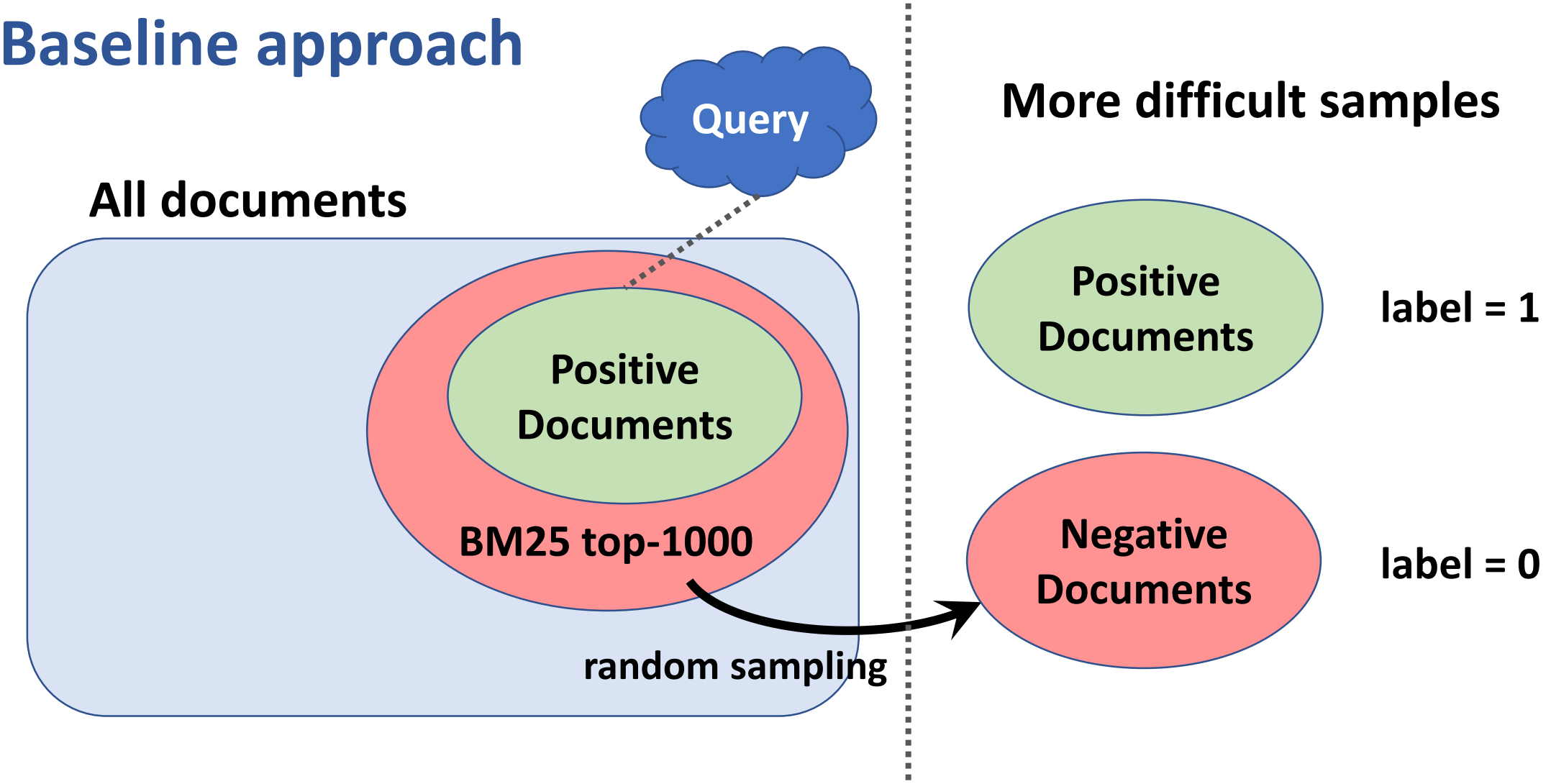
  e.g. Positive/negative documents should look similar

- **Positive/negative documents are independently trained**

  BERT cannot learn by comparing between pos./neg. documents

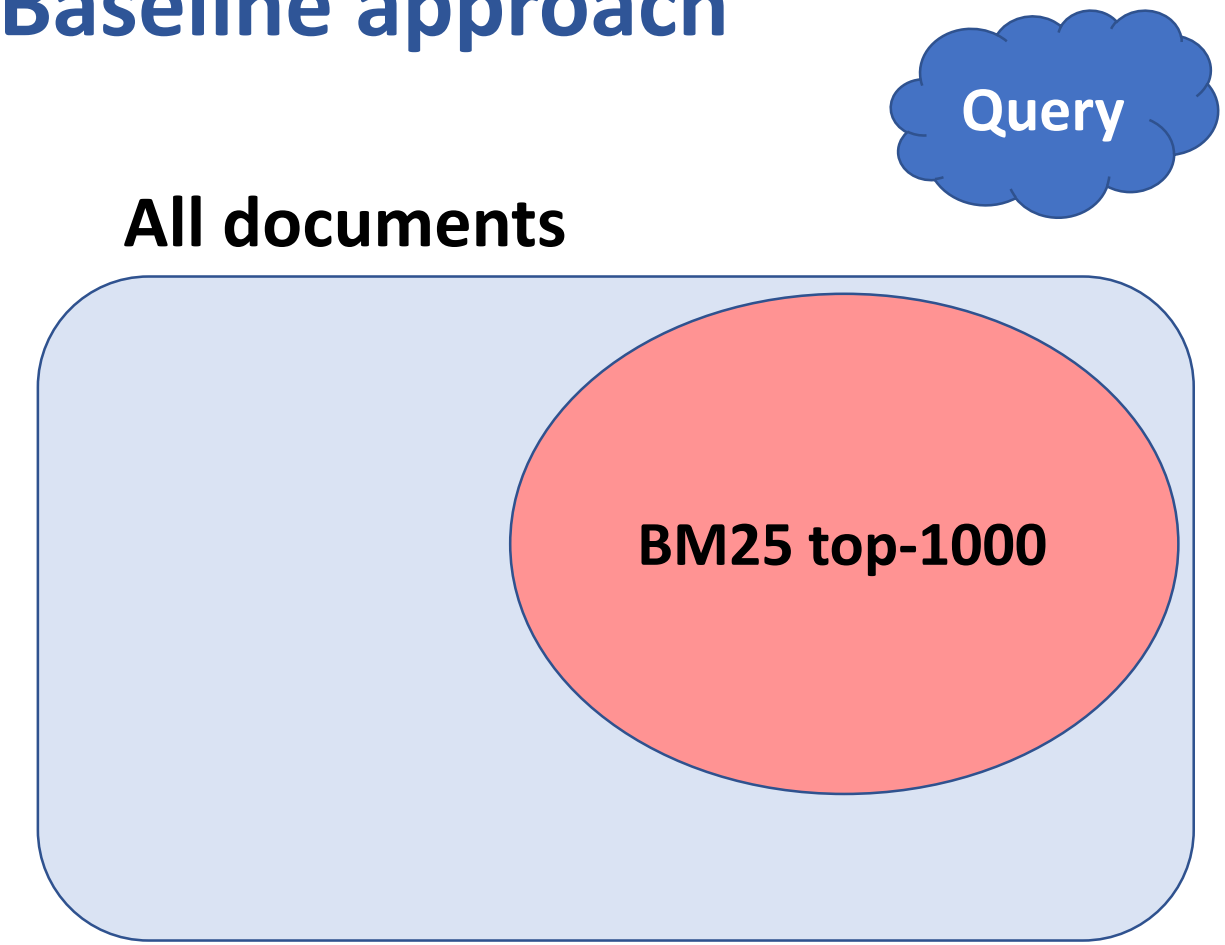
Homework 6 - Baseline

✓ **Baseline approach**



Homework 6 - Baseline

✓ **Baseline approach**



(At inference stage)

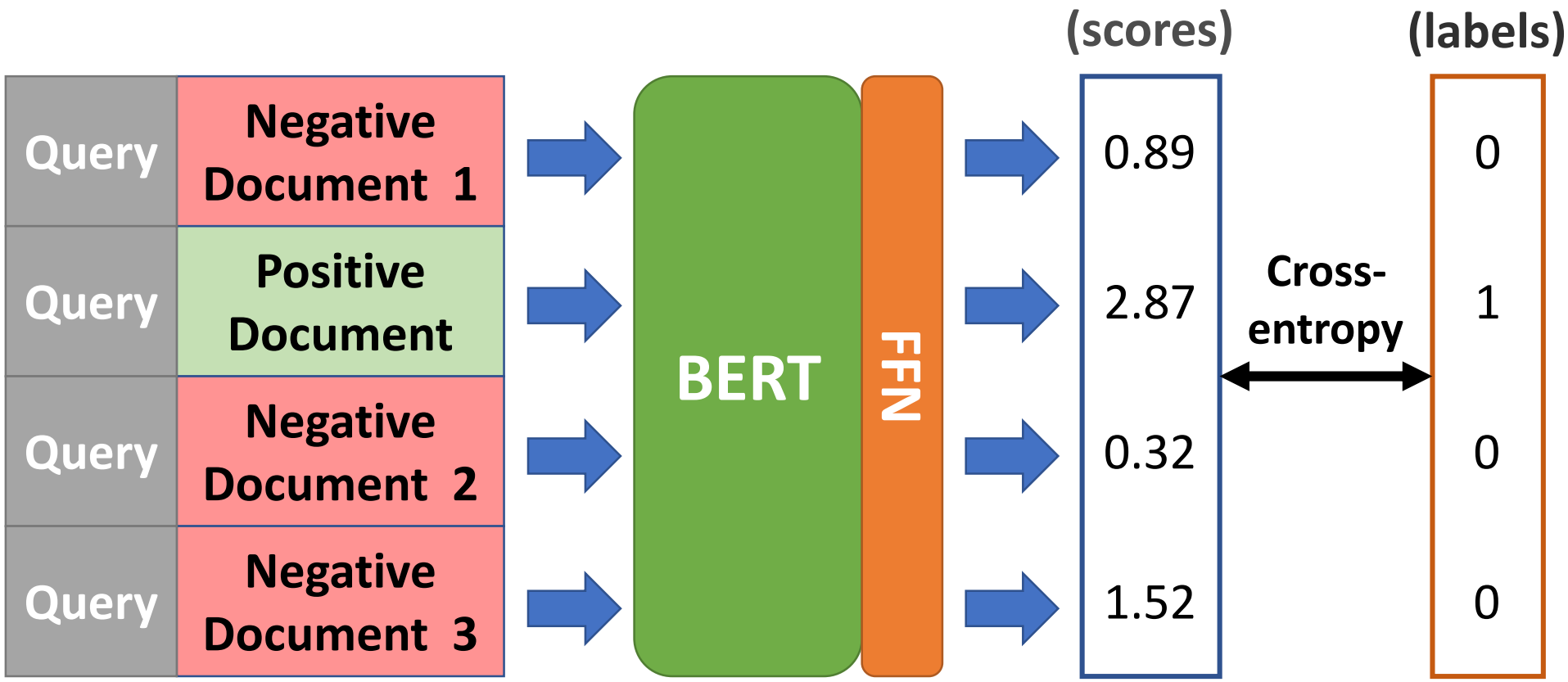
Predict relevance scores of BM25 top-1000 documents only

100x faster than naïve method :-)

Homework 6 - Baseline

✓ **Baseline approach**

Multiple-choice classification task



* Scores are independently predicted *

Homework 6 - Baseline

✓ Baseline approach

- **For each query,**

  rescore BM25 top-1000 documents with:

$$score_{new} = score_{\text{BM25}} + \alpha \cdot score_{\text{BERT}}$$

Homework 6 - Baseline

✓ Baseline settings

It takes 1~1.5 hrs to run everything on a free Kaggle kernel :-)

Hyperparameters for BERT:

- **Pretrained paramters:** "bert-base-uncased"
- **Optimizer:** AdamW w/ learning rate = 3e-5
- **Num. epochs** = 1
- **Num. of negative documents =** 3
- **Batch size** = 2
- Split 20% of training queries to grid search optimal α for BERT

Homework 6

✓ Baseline performance

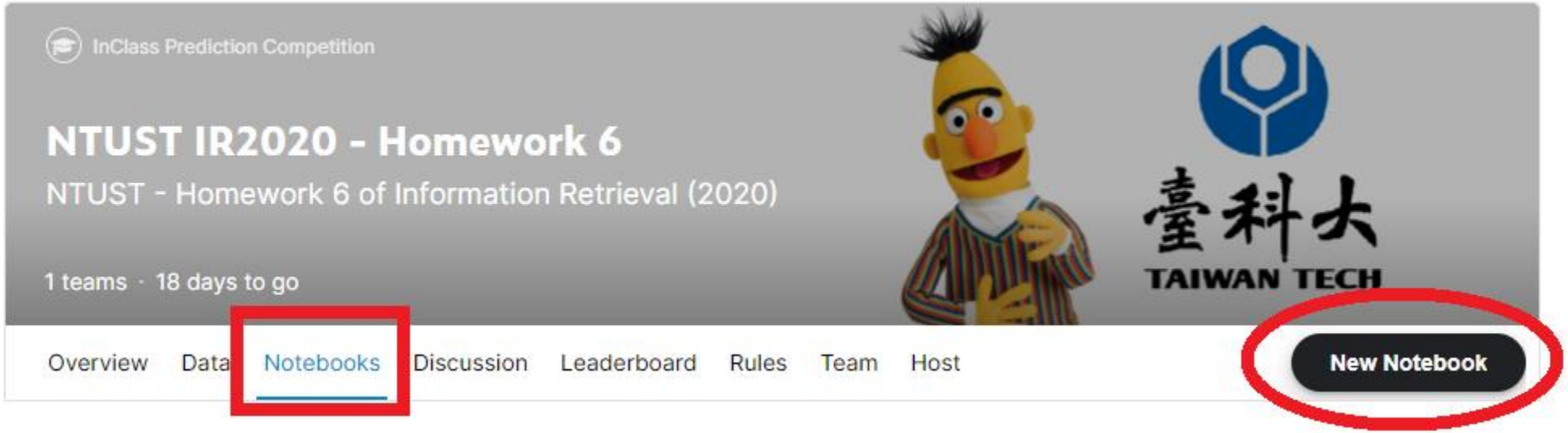
(HW6 Baseline)

Method	MAP@1000
BM25 + $\alpha \cdot$ BERT	**45.084**
BM25	39.136
Rerank BM25 w/ BERT only	30.248
Random Documents	0.007

Homework 6

✓ Free Kaggle GPU & TPU v3-8

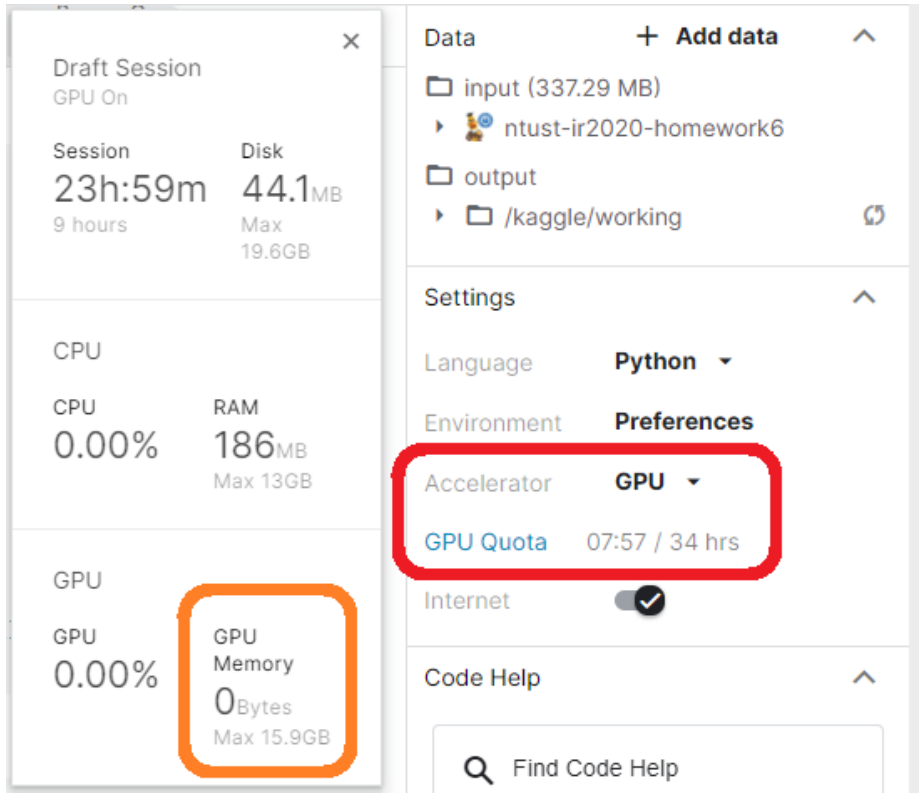
- Kaggle provides free GPU (16GB VRAM) quota of 30+ hrs/week
- Requires phone verification



Homework 6

✓ Free Kaggle GPU & TPU v3-8

- Kaggle provides free GPU (16GB VRAM) quota of 30+ hrs/week
- Requires phone verification



Homework 6

✓ Preferred tools & tutorials

- **PyTorch:**

  Official tutorial (~ 60 minutes)

  https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html

- **Huggingface's Transformers**

  Quick tour on Github & usage examples in documentation

  https://github.com/huggingface/transformers
  https://huggingface.co/transformers/

Homework 6

✓ Submission

- Kaggle URL: https://www.kaggle.com/t/4a26f9f4ba1b4feb952d5aafd98eee94
- You can submit 5 times per day.
- You can select 2 submissions to be used for your final score.
- **Deadline: 2021/1/4 23:59 (Monday)**

Homework 6

✓ Grading 15 points in total

1. Outperform baseline	get 5 points
2. Experiment report	2 point at max
3. Peer competition	8 point at max

$$\textbf{Peer score} = 8 \times \frac{\text{your MAP} - \text{baseline MAP}}{\text{1st MAP} - \text{baseline MAP}}$$

* Based on MAP of **private leaderboard** *

Thank you for your attention!

Any questions or comments?

臺科大
TAIWAN TECH

郭家廷 (Chia-Chih Kuo)

Natural Language Processing Laboratory

National Taiwan University of Science and Technology