ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

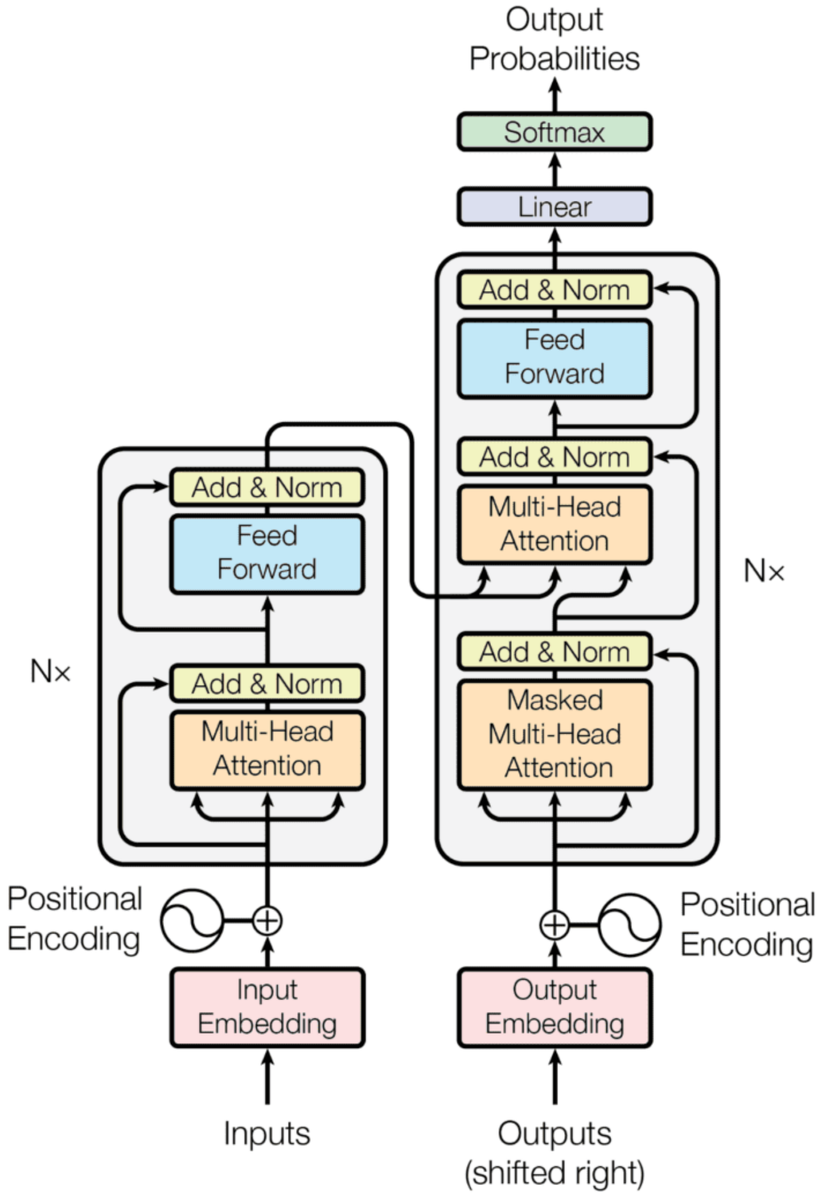
CENG 463: Introduction to Natural Language Processing
Large Language Models (Encoder Models)

Asst. Prof. Çağrı Toraman
Computer Engineering Department
ctoraman@ceng.metu.edu.tr

09.12.2025

Transformer

ENCODER



DECODER

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, NeurIPS 2017

Three types of transformer layer

Encoder (BERT)

Decoder (GPT)

Encoder-decoder (T5)

Encoder-based Models

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Released in 2018/10, NAACL 2019 best paper

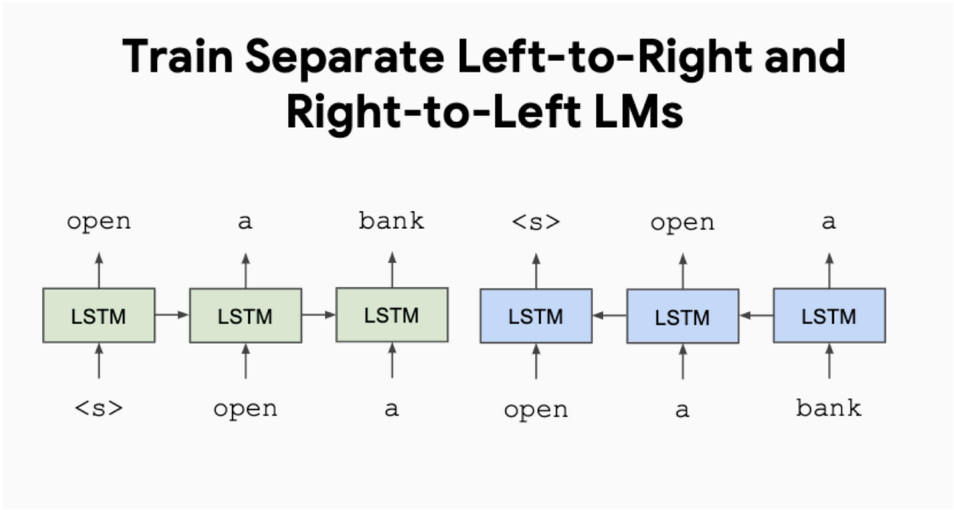

CTW
SESAME STREET®

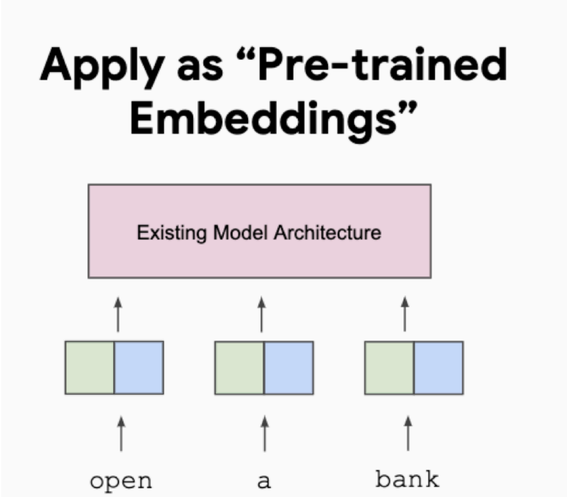
Encoder-based Models (ELMo)

Prior work: ELMo 

ELMo (Peters et al., 2018; NAACL 2018 best paper)

- Train two separate **unidirectional** LMs (left-to-right and right-to-left) based on **LSTMs**
- **Feature-based** approach: pre-trained representations used as input to task-specific models
- Trained on **single sentences** from 1B word benchmark (Chelba et al., 2014)





Encoder-based Models (BERT*)

Learn representations based on **bidirectional context**

Both left and right contexts are important to understand the meaning of words

Example #1: we went to the river bank
Example #2: I need to go to bank to make a deposit

SOTA performance on **sentence-level** and **token-level** tasks

**Some slides are adapted from Princeton University COS597G*

Encoder-based Models (BERT)

Pre-training objectives:

Masked language modeling

Next sentence prediction

Encoder-based Models (BERT)

Masked Language Modeling (MLM)

- Q: Why we can't do language modeling with bidirectional models?

"cheat" by looking at the future words

- Solution: Mask out k% of the input words, and then predict the masked words

store gallon

the man went to [MASK] to buy a [MASK] of milk

Encoder-based Models (BERT)

What is the value of k?

$k = 15\%$

Why?

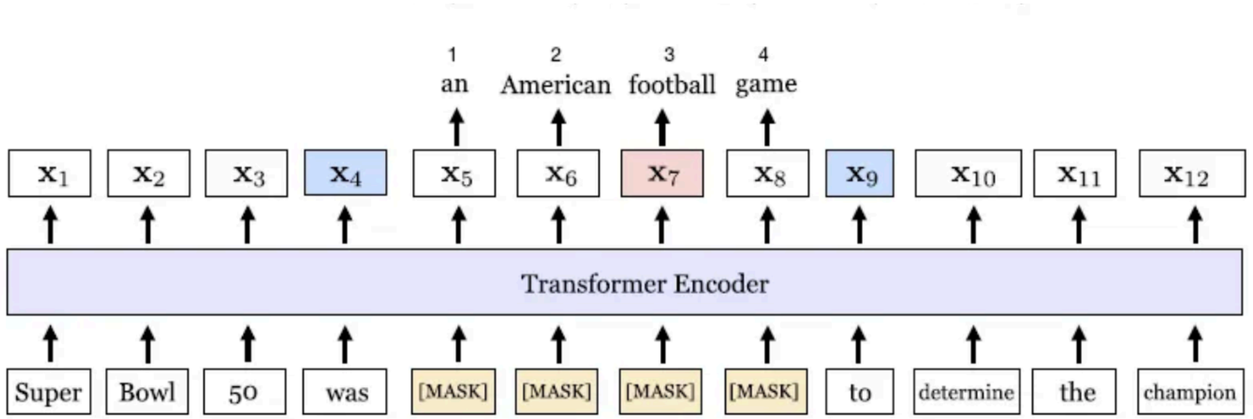
Encoder-based Models (BERT)

How are masked tokens selected?

15% tokens are uniformly sampled

<u>Alternative masking solutions:</u>

SpanBERT



Encoder-based Models (BERT)

For the 15% predicted words:

80% of the time, they replace it with [MASK] token

went to the store ⟶ went to the [MASK]

10% of the time, they replace it with a random word in the vocabulary

went to the store ⟶ went to the running

10% of the time, they keep it unchanged

went to the store ⟶ went to the store

Why?

Encoder-based Models (BERT)

Next Sentence Prediction (NSP)

- Motivation: many NLP downstream tasks require understanding the relationship between two sentences (natural language inference, paraphrase detection, QA)
- NSP is designed to reduce the gap between pre-training and fine-tuning

[CLS]: a special token always at the beginning

[SEP]: a special token used to separate two segments

Input = `[CLS] the man went to [MASK] store [SEP]`

`he bought a gallon [MASK] milk [SEP]`

Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP]`

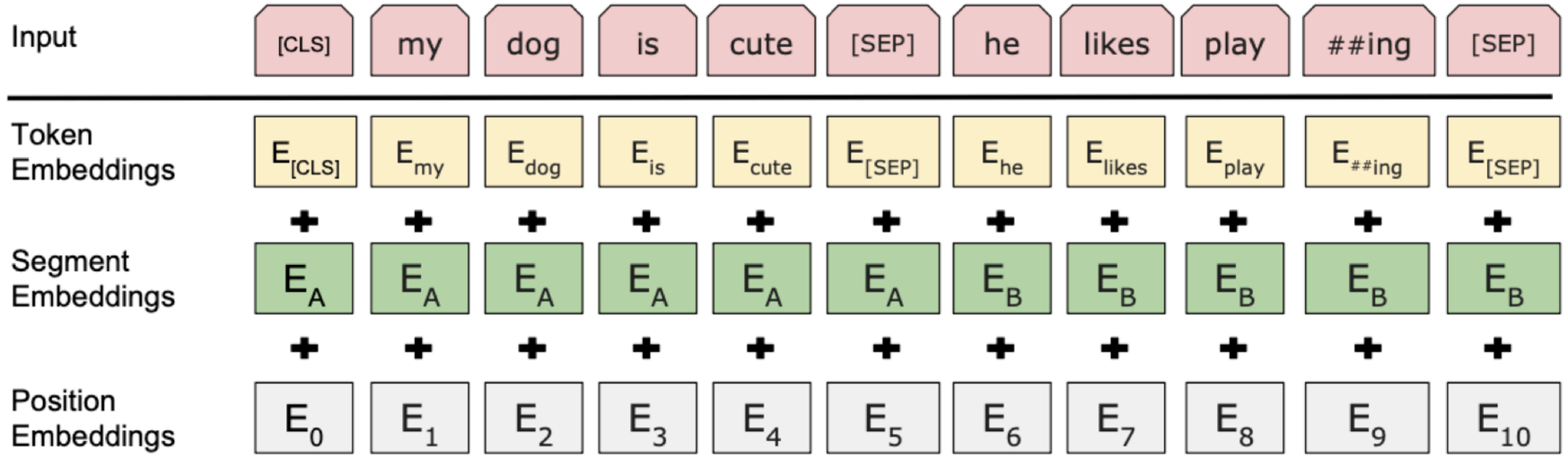
`penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

They sample two contiguous segments for 50% of the time and another random segment from the corpus for 50% of the time

Encoder-based Models (BERT)

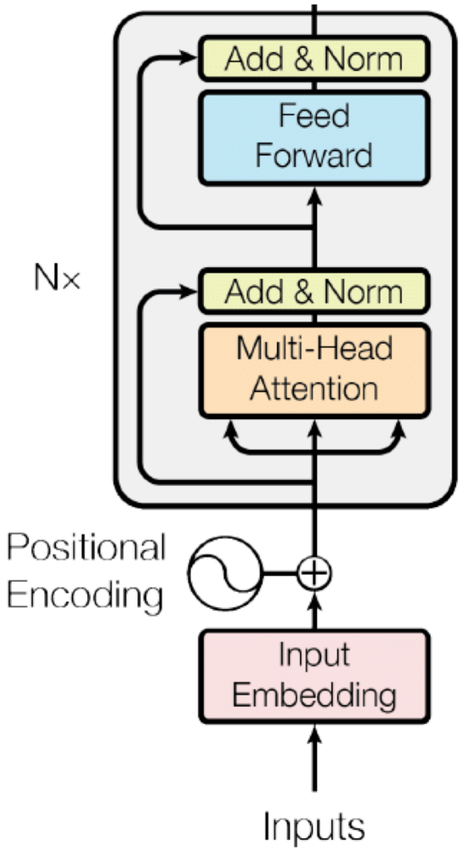
Input embeddings:



Vocabulary size: 30k tokens (by WordPiece)

Encoder-based Models (BERT)

Transformer:
Encoder Layers: 6
FFNN Hidden Layer Units: 512
Attention Heads: 8



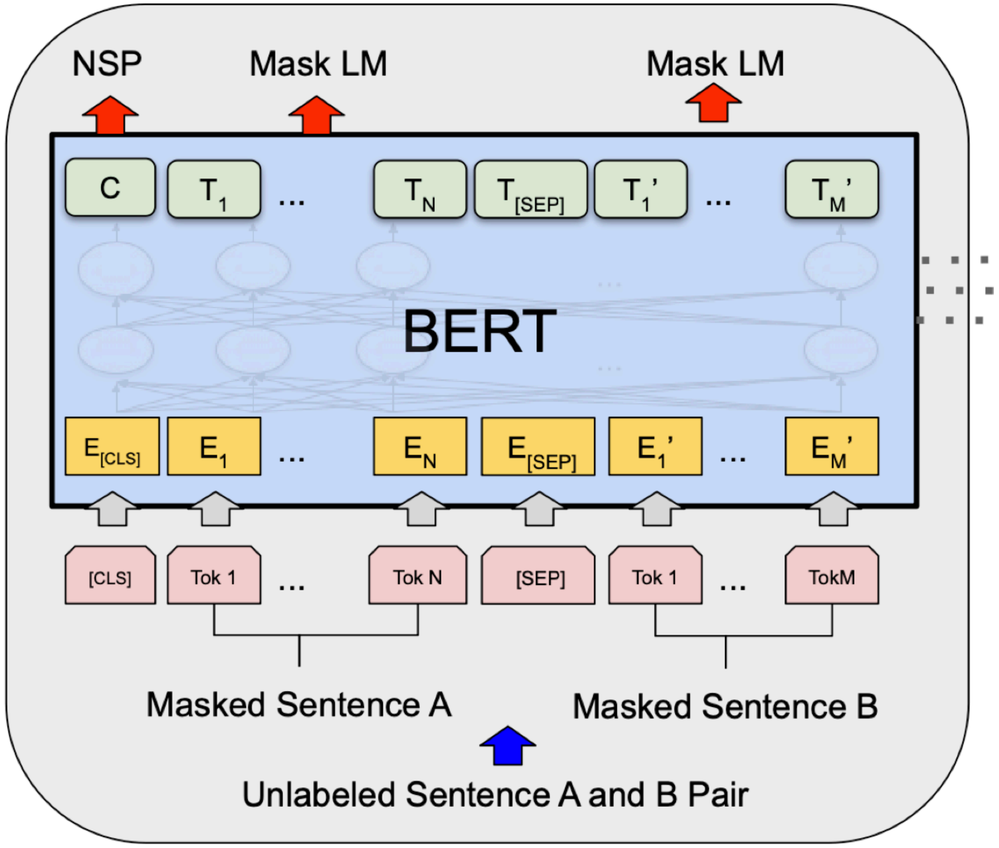
- BERT-base: 12 layers, 768 hidden size, 12 attention heads, 110M parameters
- BERT-large: 24 layers, 1024 hidden size, 16 attention heads, 340M parameters

Same as OpenAI GPT

OpenAI GPT was trained on BooksCorpus only!

- Training corpus: Wikipedia (2.5B) + BooksCorpus (0.8B)
- Max sequence size: 512 word pieces (roughly 256 and 256 for two non-contiguous sequences)
- Trained for 1M steps, batch size 128k

Encoder-based Models (BERT)



Pre-training

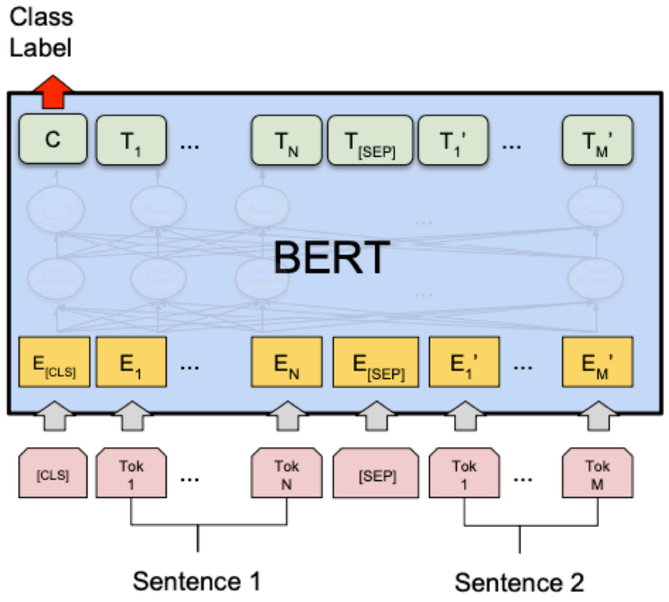
- MLM and NSP are trained together
- [CLS] is pre-trained for NSP
- Other token representations are trained for MLM

Encoder-based Models (BERT)

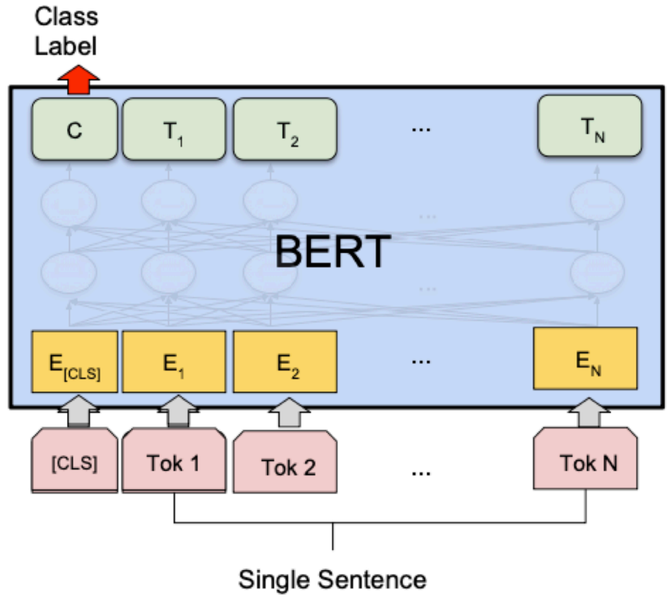
Fine-tuning BERT

"Pretrain once, finetune many times."

sentence-level tasks



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA

Encoder-based Models (BERT)

Sentence-level tasks

- Sentence pair classification tasks:

MNLI

Premise: A soccer game with multiple males playing.

Hypothesis: Some men are playing a sport.

{entailment, contradiction, neutral}

QQP

Q1: Where can I learn to invest in stocks?

Q2: How can I learn more about stocks?

{duplicate, not duplicate}

- Single sentence classification tasks:

SST2

rich veins of funny stuff in this movie

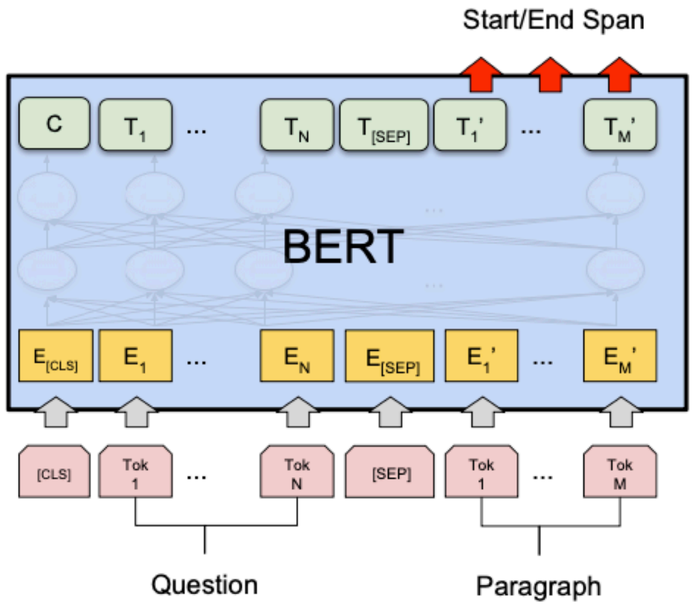
{positive, negative}

Encoder-based Models (BERT)

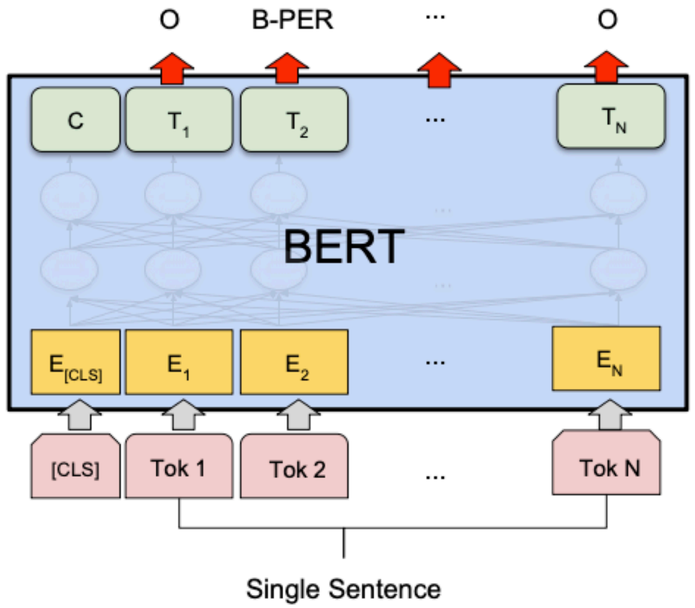
Fine-tuning BERT

"Pretrain once, finetune many times."

token-level tasks



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Encoder-based Models (BERT)

Token-level tasks

- Extractive question answering e.g., SQuAD (Rajpurkar et al., 2016)

SQuAD

Question: The New York Giants and the New York Jets play at which stadium in NYC ?

Context: The city is represented in the National Football League by the New York Giants and the New York Jets , although both teams play their home games at MetLife Stadium in nearby East Rutherford , New Jersey , which hosted Super Bowl XLVIII in 2014 .

(Training example 29,883)

MetLife Stadium

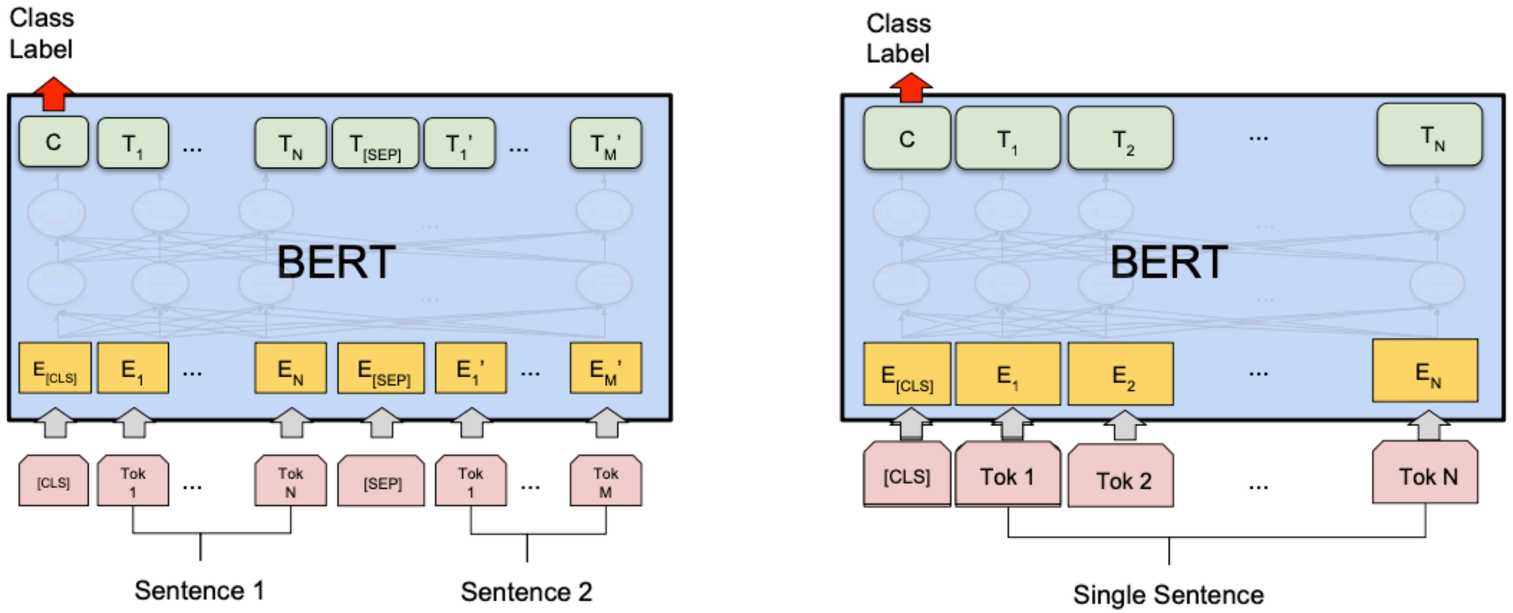
- Named entity recognition (Tjong Kim Sang and De Meulder, 2003)

CoNLL 2003 NER

John	Smith	lives	in	New	York
B-PER	I-PER	O	O	B-LOC	I-LOC

Encoder-based Models (BERT)

Fine-tuning BERT

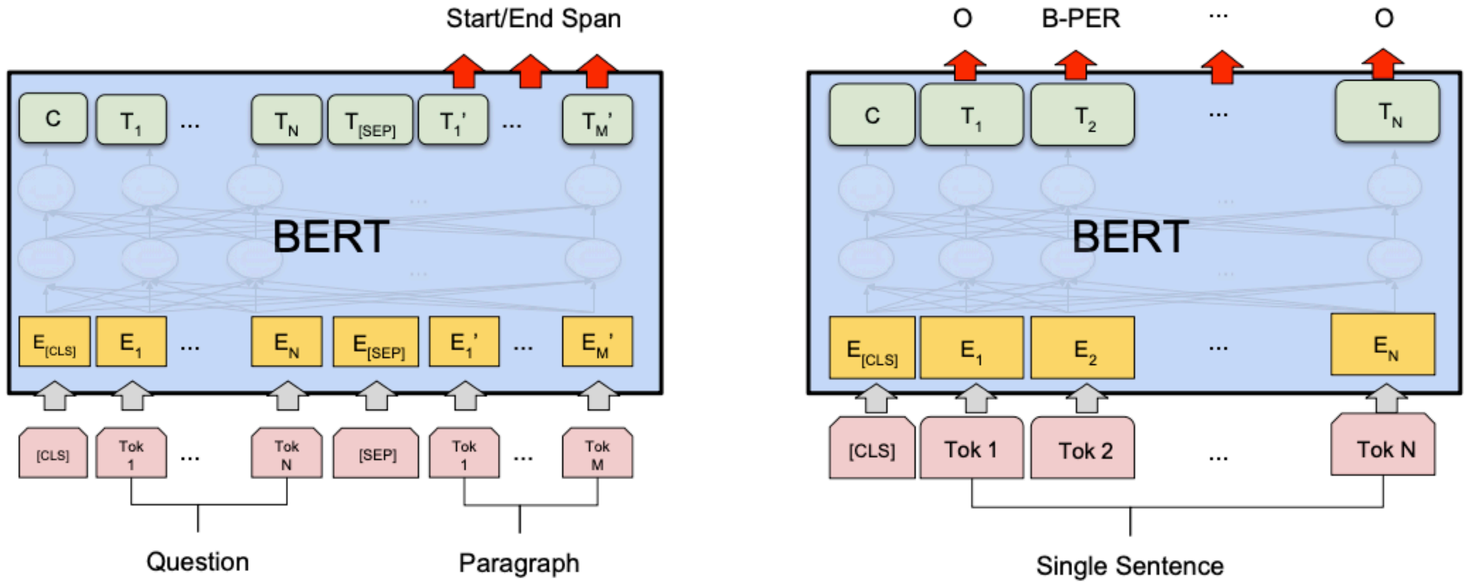


- For sentence pair tasks, use [SEP] to separate the two segments with segment embeddings
- Add a linear classifier on top of [CLS] representation and introduce $C \times h$ new parameters

C: # of classes, h: hidden size

Encoder-based Models (BERT)

Fine-tuning BERT



- For token-level prediction tasks, add linear classifier on top of hidden representations

Q: How many new parameters?

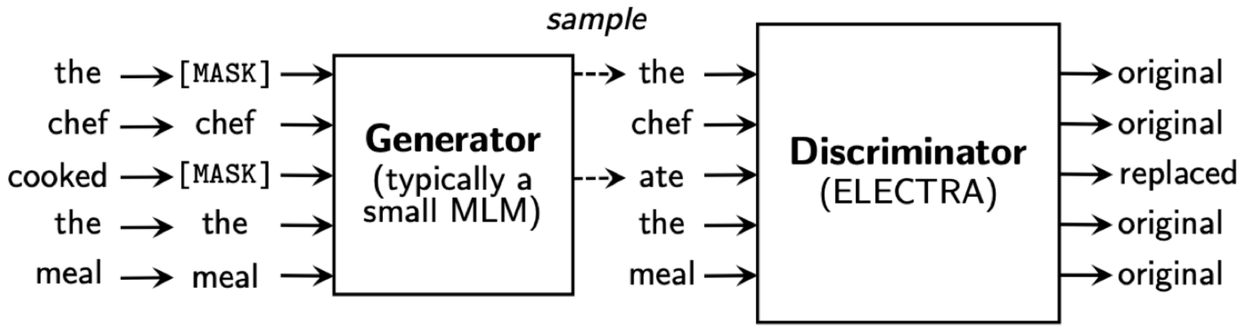
Encoder-based Models (BERT)

BERT era (2018-2021):

Liu, Y. (2019). **Roberta**: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, *364*.
Trained on 10x data & longer, and removing Next Sentence Prediction

Lan, Z. (2019). **Albert**: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
Increasing model sizes by sharing model parameters across layers

Clark, K. (2020). **Electra**: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
Providing a more efficient training method by predicting 100% of tokens instead of 15% of tokens



Encoder-based Models (BERT)

BERT era (2018-2021):

Sanh, V. (2019). **DistilBERT**, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
Making BERT smaller in terms of model parameters

Conneau et al., (2020). “Unsupervised **Cross-lingual** Representation Learning at Scale,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451.
Multilingual learning by training single model on 104 languages and sharing ~100k vocabulary

Beltagy, I., Peters, M. E., & Cohan, A. (2020). **Longformer**: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
Models that handle long contexts (>512 tokens)

Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021, January). **Finbert**: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 4513-4519).
Models adapted to specific domains

Encoder-based Models (BERT)

After BERT era (modernization efforts after 2021):

MosaicBERT
Portes, J., Trott, A., Havens, S., King, D., Venigalla, A., Nadeem, M., ... & Frankle, J. (2023). MosaicBERT: A bidirectional encoder optimized for fast pretraining. *Advances in Neural Information Processing Systems*, *36*, 3106-3130.

Chicago

ModernBERT
Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., ... & Poli, I. (2025, July). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2526-2547).

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

Thanks for your participation!

Çağrı Toraman
09.12.2025