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

강동규

DeepSync, South Korea

Problem – Using Labeled Data

Labeled Data

Cost

Small amount of Data

Solution of data scarcity - GPT

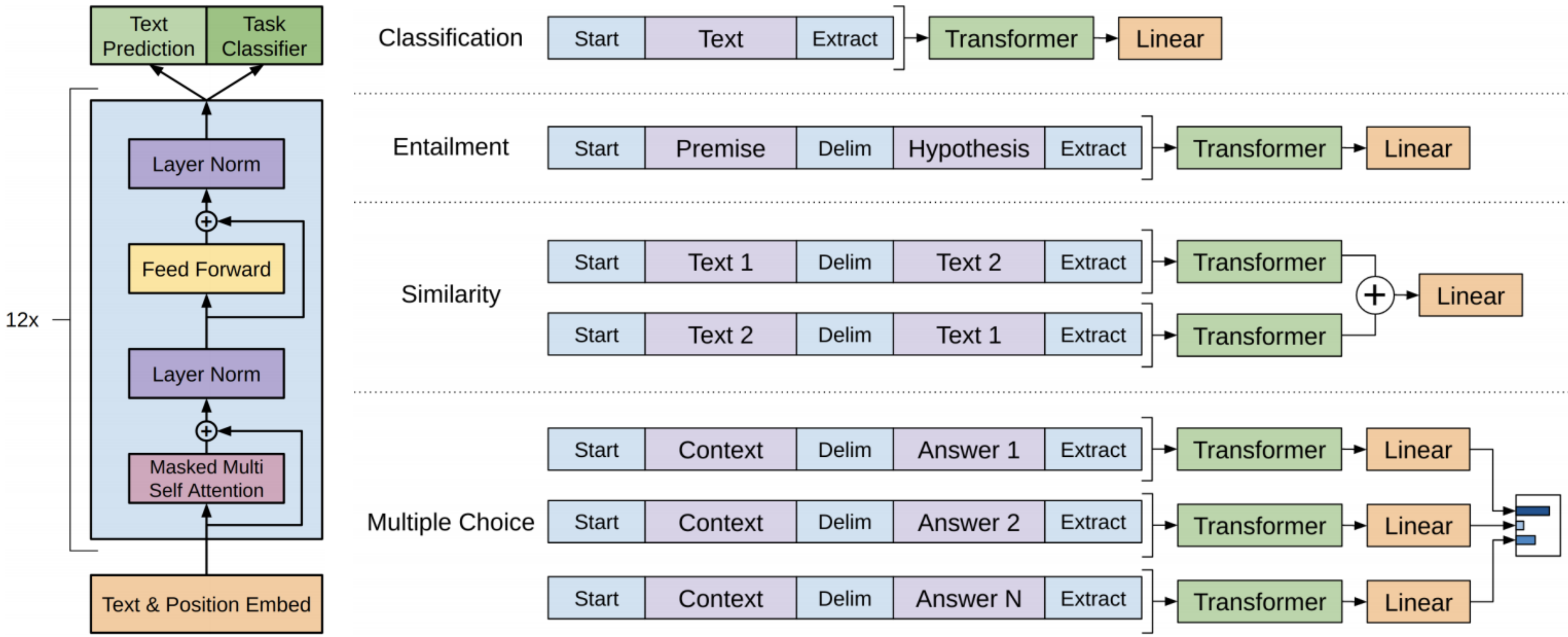


Fig1

Solution of data scarcity - GPT

Deep Sync

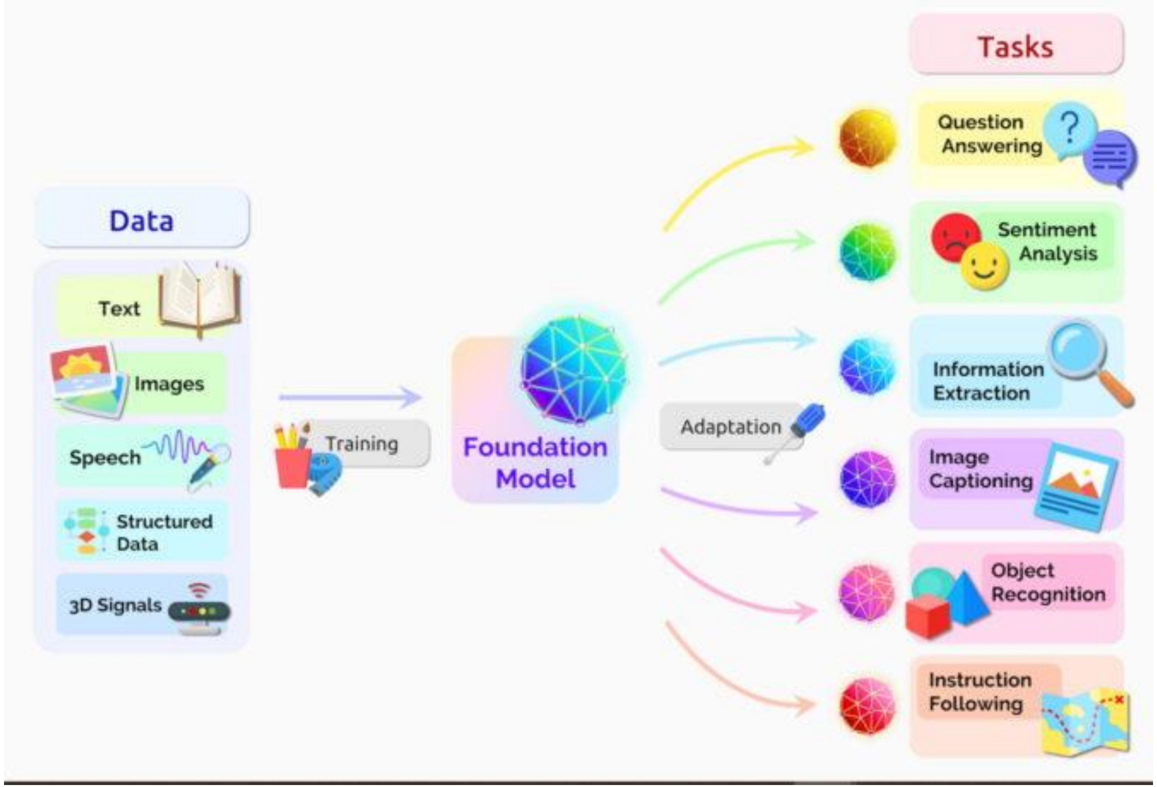
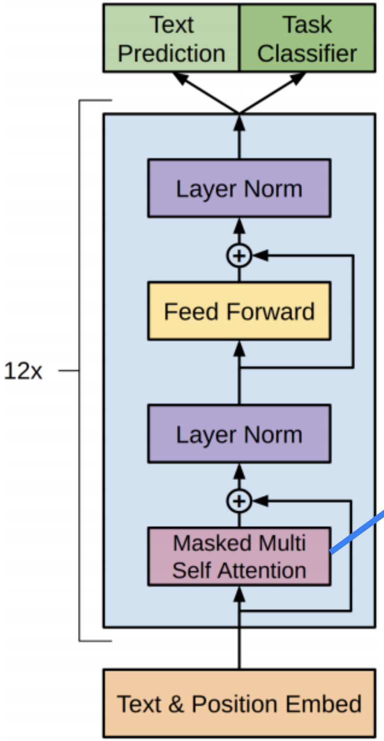


Fig2

Problem of GPT



	I	am	a	boy	pad	pad
I						
am						
a						
boy						
pad						
pad						

Problem of GPT

Deep Sync

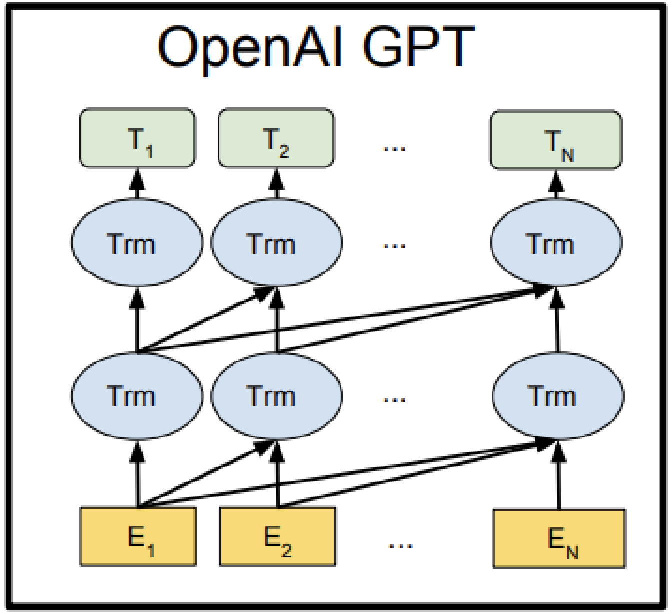
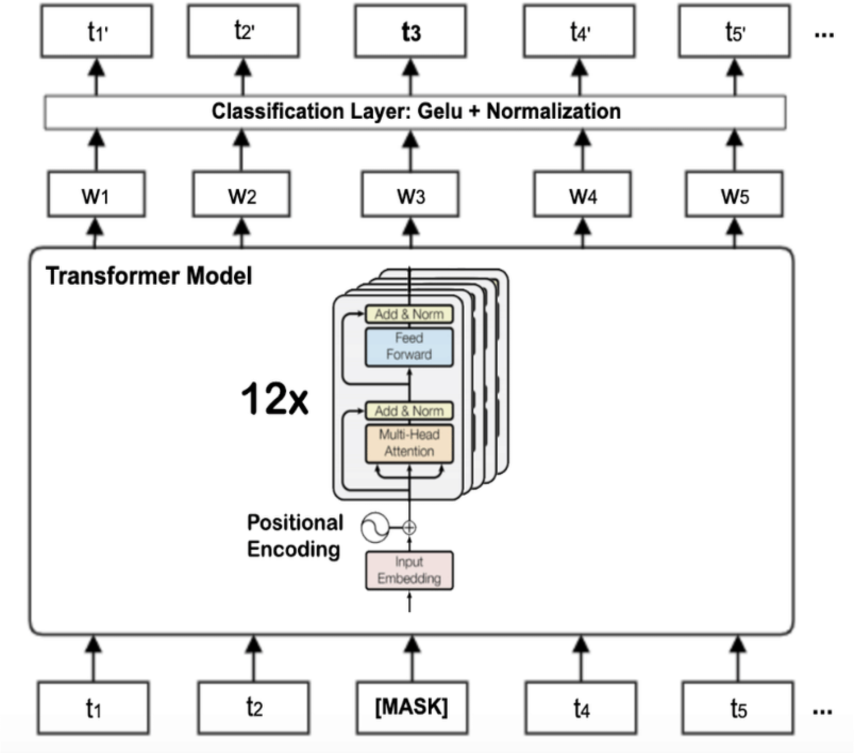


Fig4

Unidirectional context learning!

BERT



t1'
t2'
t3
t4'
t5'
...
Classification Layer: Gelu + Normalization
W1
W2
W3
W4
W5
Transformer Model
12x
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
t1
t2
[MASK]
t4
t5
...

Fig5

BERT

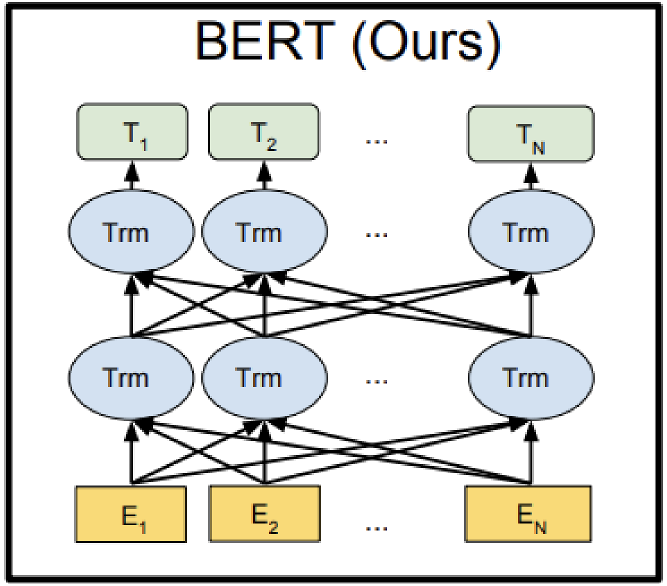


Fig6

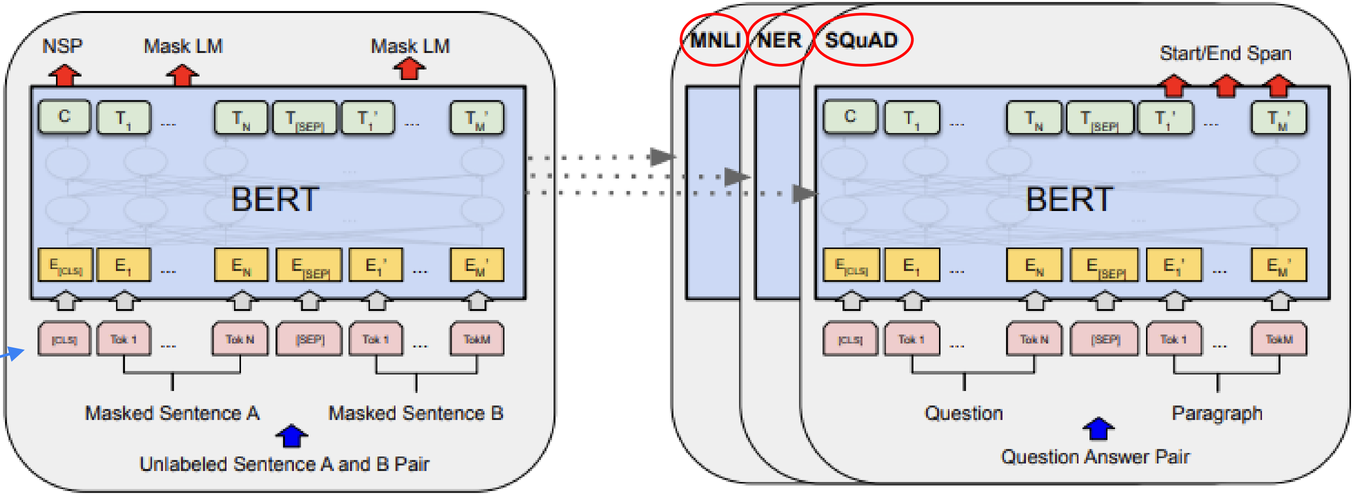
- Bidirectional context learning

→ Learning both left and right context

→ Learning more general feature

BERT

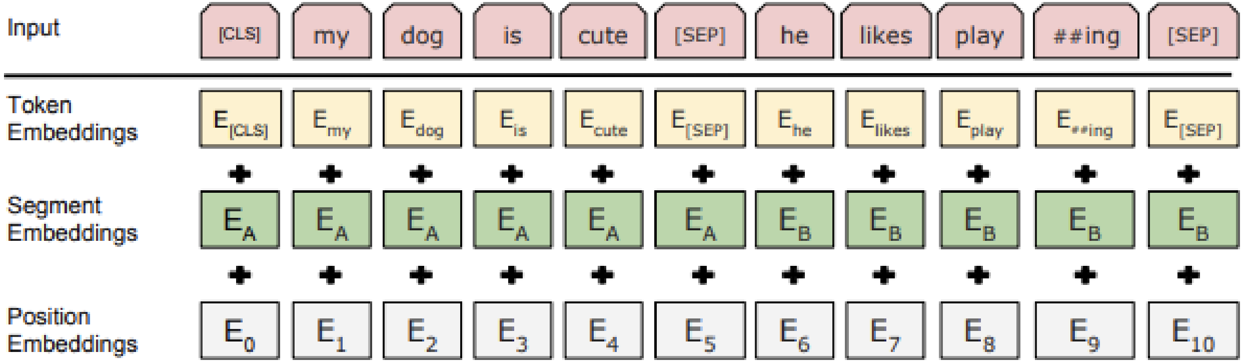
Unlabeled Data



Pre-training Fine-Tuning

Fig7

Implementation – Positional Embedding



Token	Word embedding token
Segment	Separate sentence
Position	Word positional embedding

CLS token	Using on classification task
SEP token	Separate sentence

Implementation – MLM

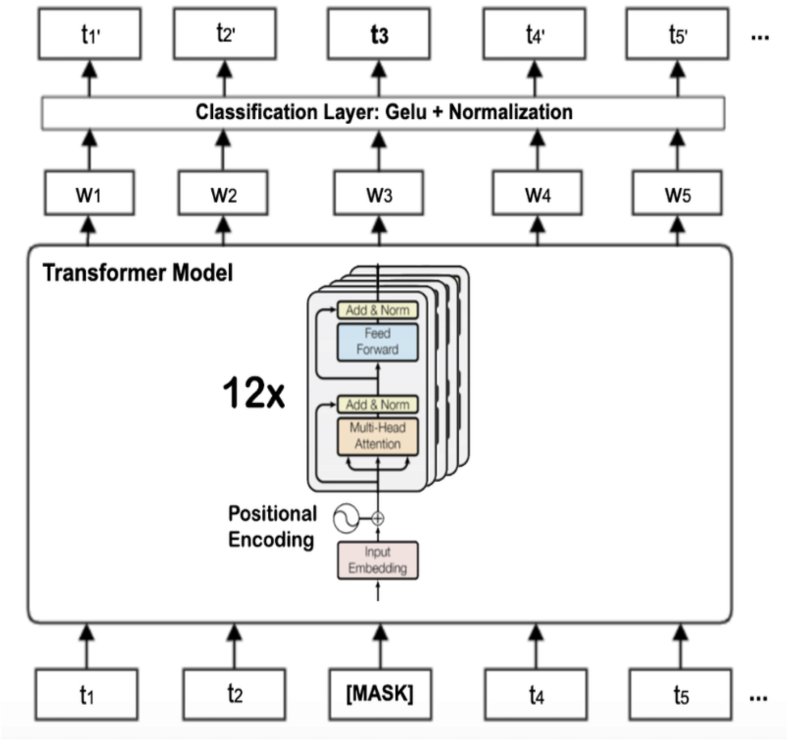


Fig5

The man went to the store and buy a snack

→ ←

The man went to the [MASK] and buy a snack

- **Masked Language Model(MLM)**

Objective	**Predict the original vocab id of masked word**
Effect	**Allow model to pretrain deep feature**

Implementation – MLM (Detail)

Deep Sync

Original	The man went to the store and buy a snack

Select 15% words randomly in original words

80% Mask	The man went to the [MASK] and buy a snack
10% Random	The man went to the cat and buy a snack
10% Equal	The man went to the store and buy a snack

Why? → Using only [MASK] creates a mismatch between pre-training and fine-tuning

Implementation – MLM (Detail)

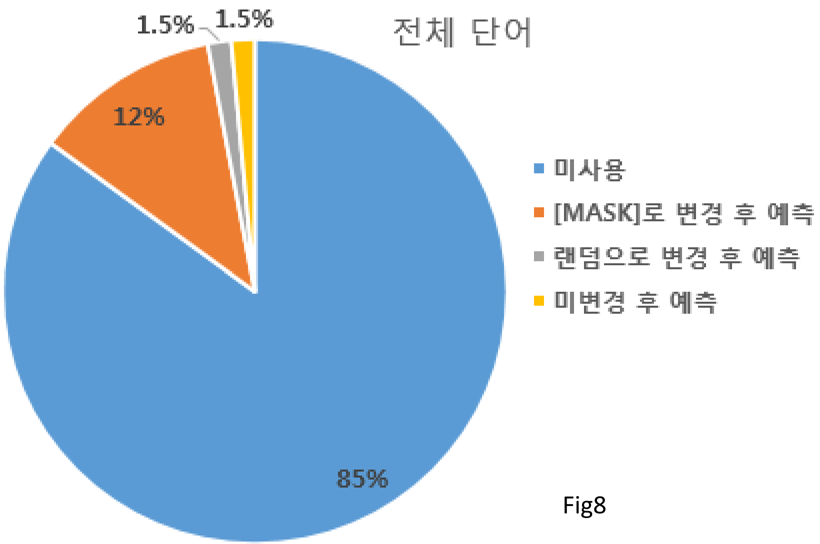
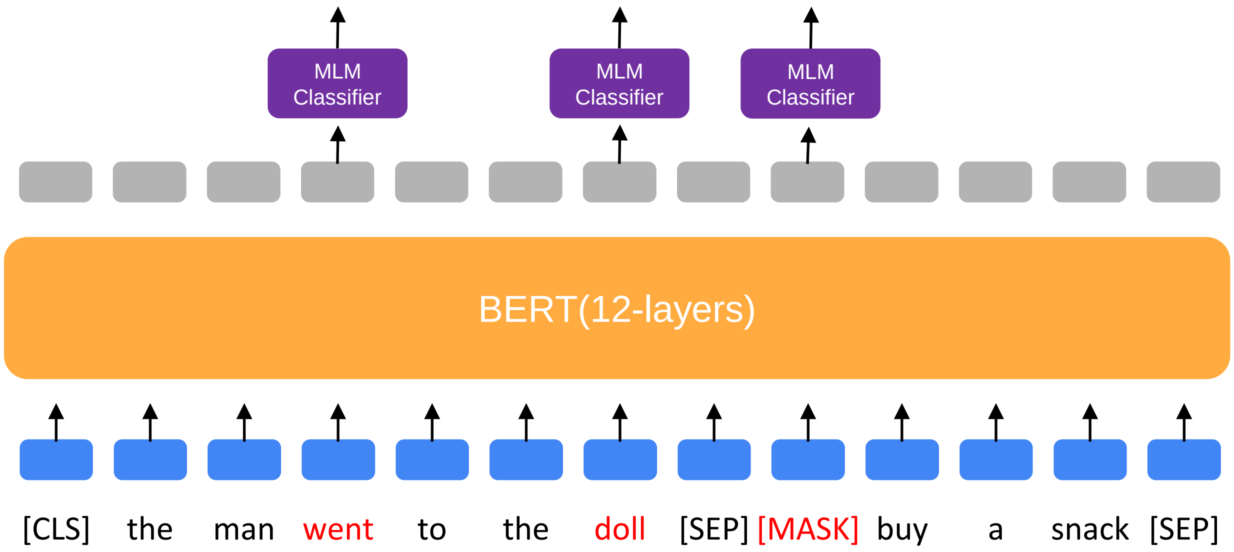


Fig8

Implementation – MLM (Detail)

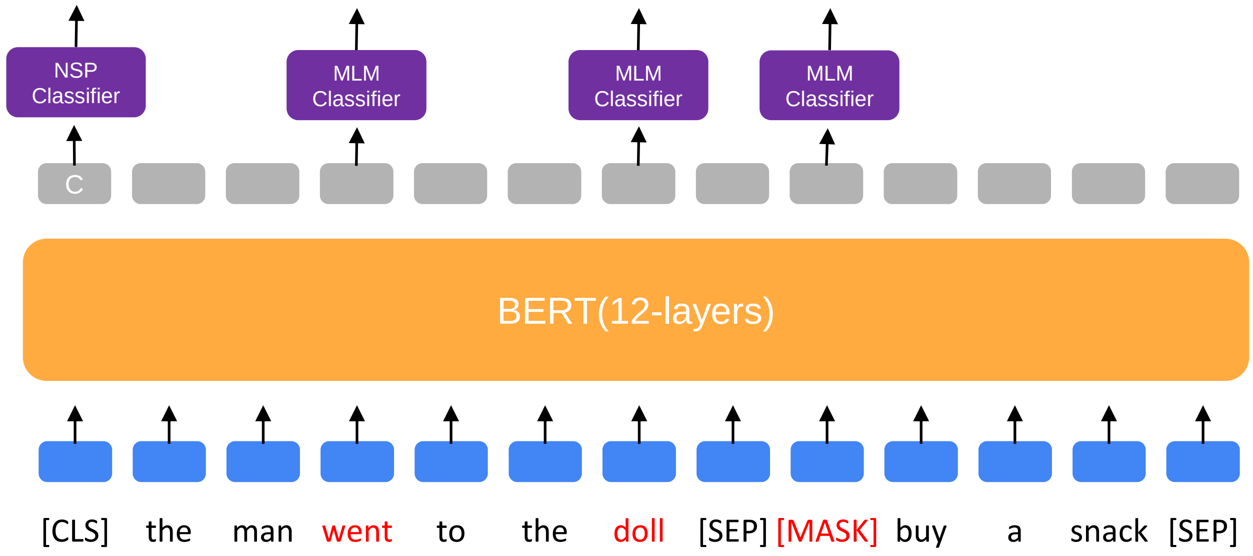


Implementation – NSP

Deep Sync

Sentence A	The man went to the store
Sentence B	He bought a gallon of milk
Label	IsNext
5:5	
Sentence A	The man went to the store
Sentence B	Dogs are so cute
Label	NotNext

Implementation – NSP



Result – GLUE

GLUE – General Language Understanding Evaluation

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Conclusion

1. BERT can learn deep feature by using bidirectional architecture

2. BERT can be fine-tuned and achieved SOTA on various NLP tasks

Reference

Fig1, Fig3 - https://paperswithcode.com/method/gpt

Fig2 - https://blogs.nvidia.co.kr2023/04/04/what-are-foundation-models/

Fig4, Fig6, Fig7 - https://arxiv.org/pdf/1810.04805.pdf

Fig5 - https://www.researchgate.net/figure/The-Transformer-based-BERT-base-architecture-with-twelve-encoder-blocks_fig2_349546860

Fig8 - https://wikidocs.net/115055

GLUE explanation - https://mccormickml.com/2019/11/05/GLUE/

Thank you for your time