신입생 Deep Learning 기초 교육

BERT & GPT

Multimodal Language Cognition Lab,
Kyungpook National University

2023.02.16

MLCL Kyungpook National University

Background

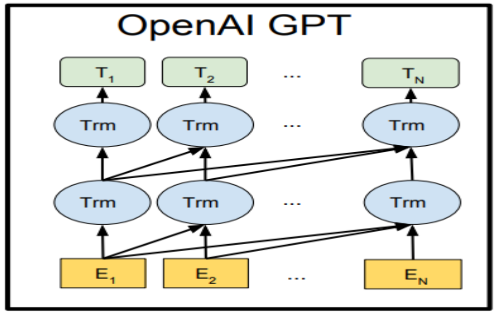
2017(June)

Transformer

2018(June)

GPT

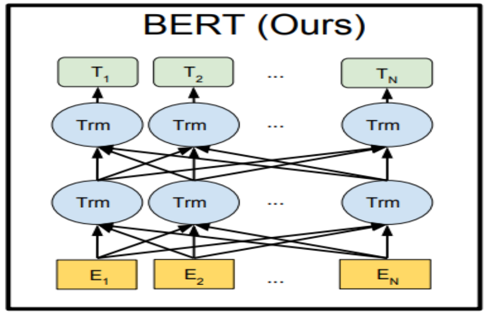
- Transformer Decoder
- Generative Model



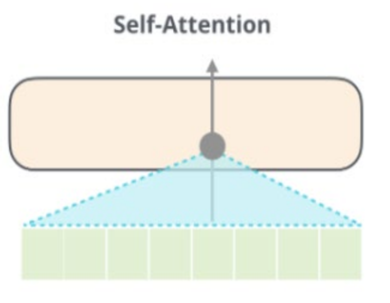
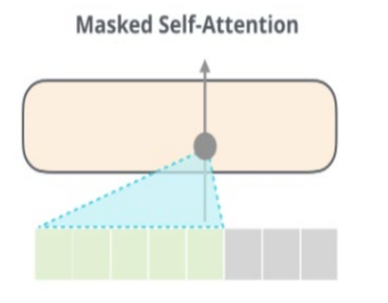
2018(Oct)

BERT

- Transformer Encoder
- Feature vector 생성



Background

	BERT		GPT-1	
Attention range	 	Consider all tokens	 	Consider previous tokens
Generation	X		O	
Fine-tuning	required		auxiliary	

Bidirectional Encoder Representations from Transformer

Pre-training

- Unlabeled date

  bookscorpus: 800M words
  english wikipedia: 2,500M words

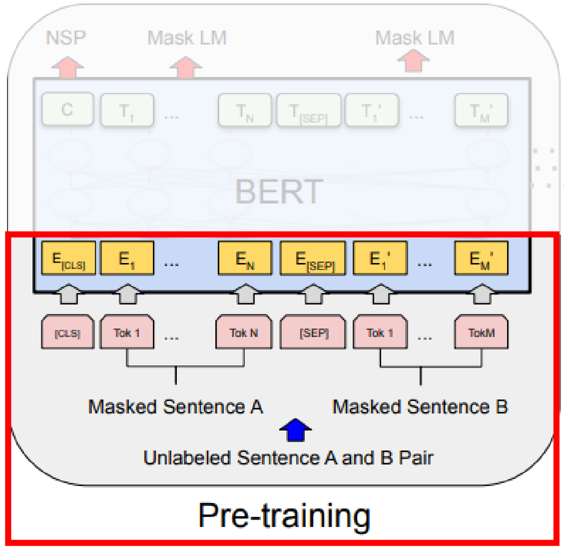
- Embedding
- **MLM** (Masked language model)
- **NSP** (Next Sentence Prediction)

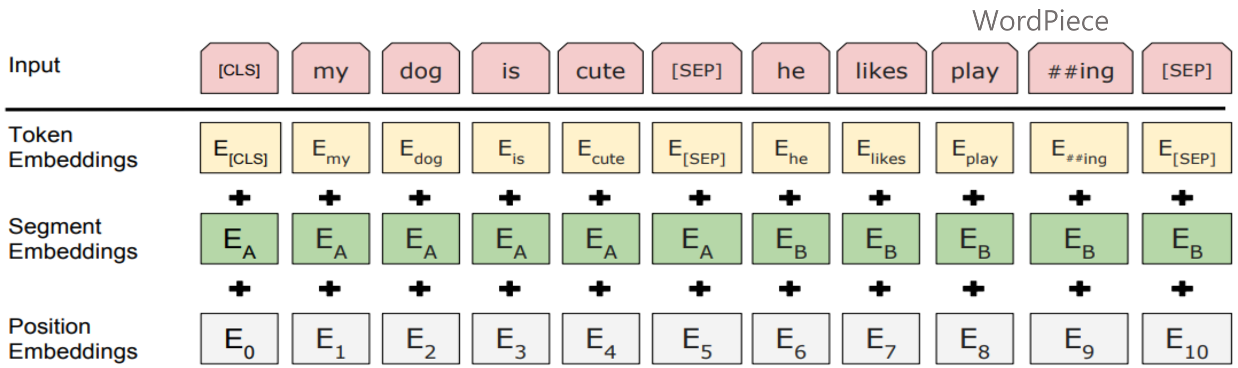
Fine-tuning

- Labeled date
- Embedding
- fine-tuning with just one additional

  output layer

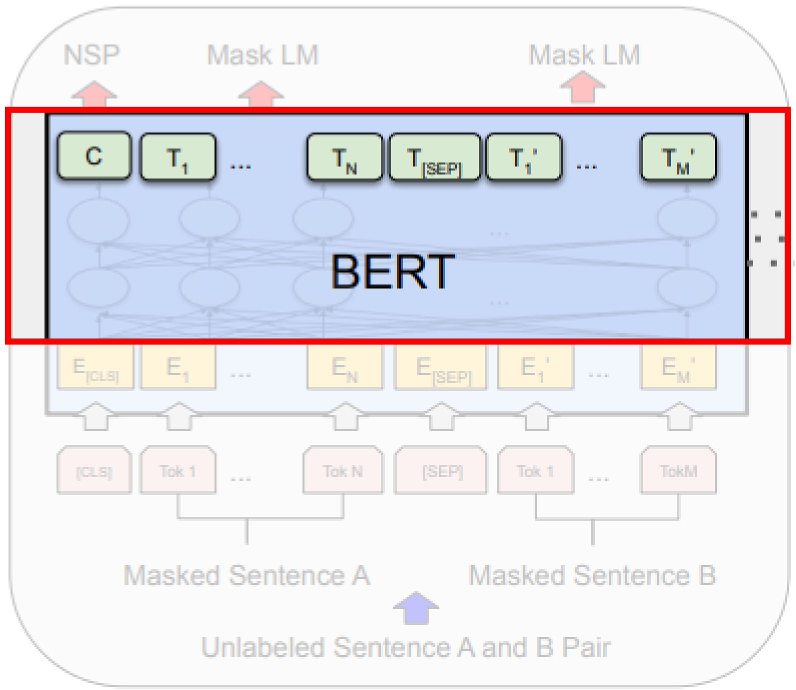
BERT: Embedding



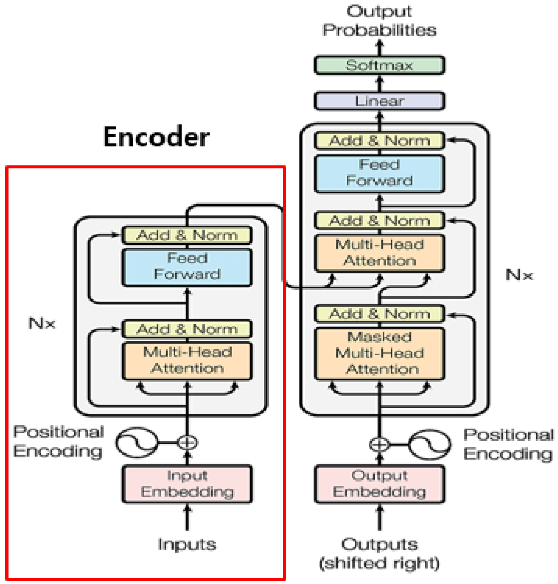


- Token embeddings: WordPiece embeddings with a 30,000 token vocabulary
- Segment embeddings: Add a different number to each sentence
- Position embeddings: Indicates the order of the sentences

BERT: Encoder

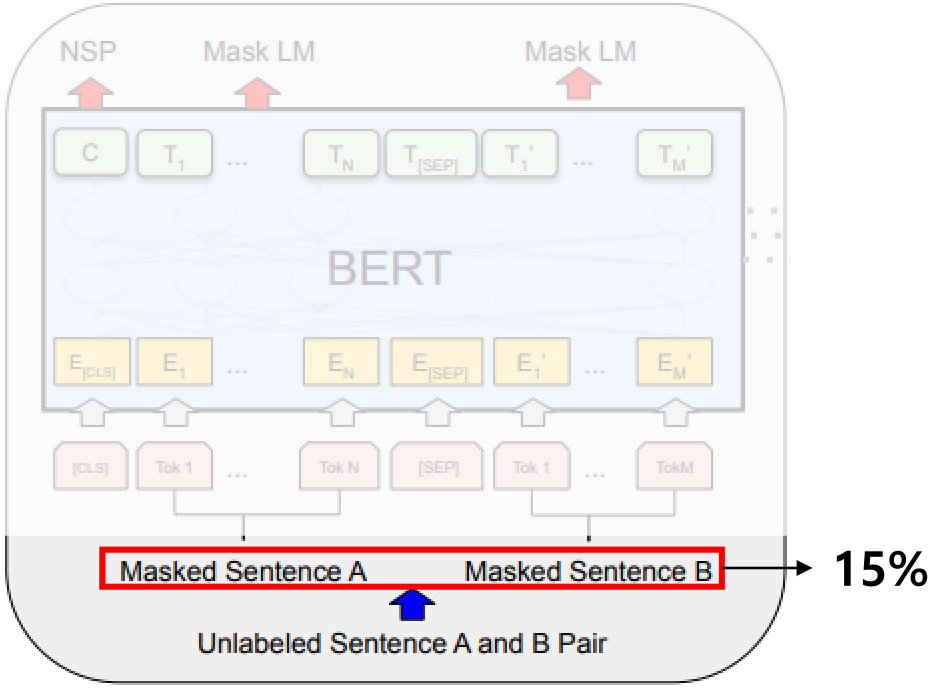


Pre-training



	BERT-Base	BERT-Large
Hidden size	768	1024
Layer	12	24
Total parameters	110M	340M

BERT: MLM(Masked Language Model)

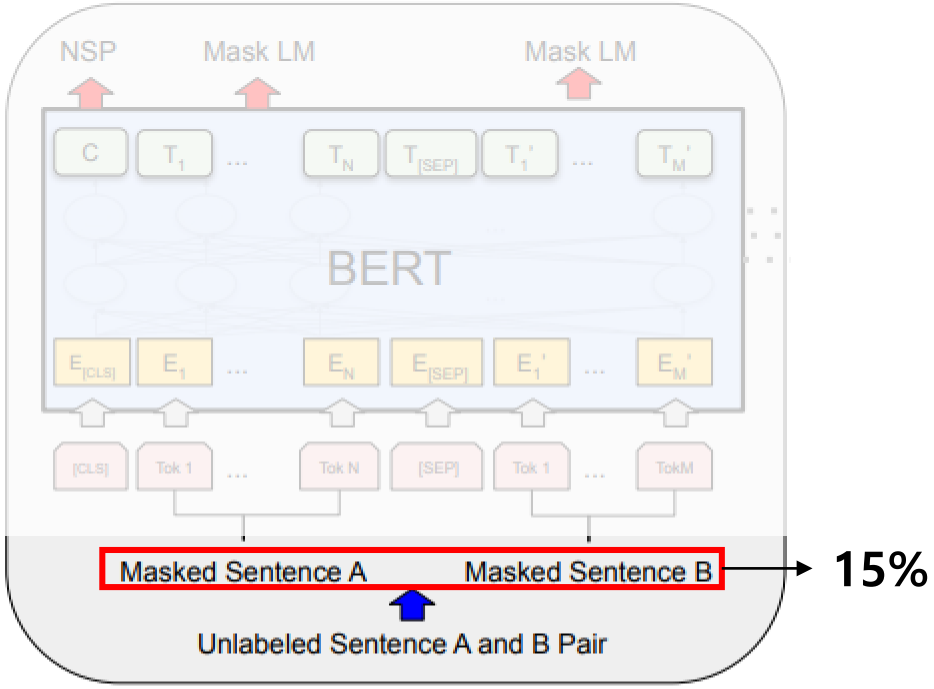


Pre-training

- **Pre-training, Fine-tuning mismatch**
- **Mask ratio**
  - 80% of the time: my dog is hairy → my dog is **[MASK]**
  - 10% of the time: my dog is hairy → my dog is **apple**
  - 10% of the time: my dog is **hairy**

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

BERT: MLM(Masked Language Model)

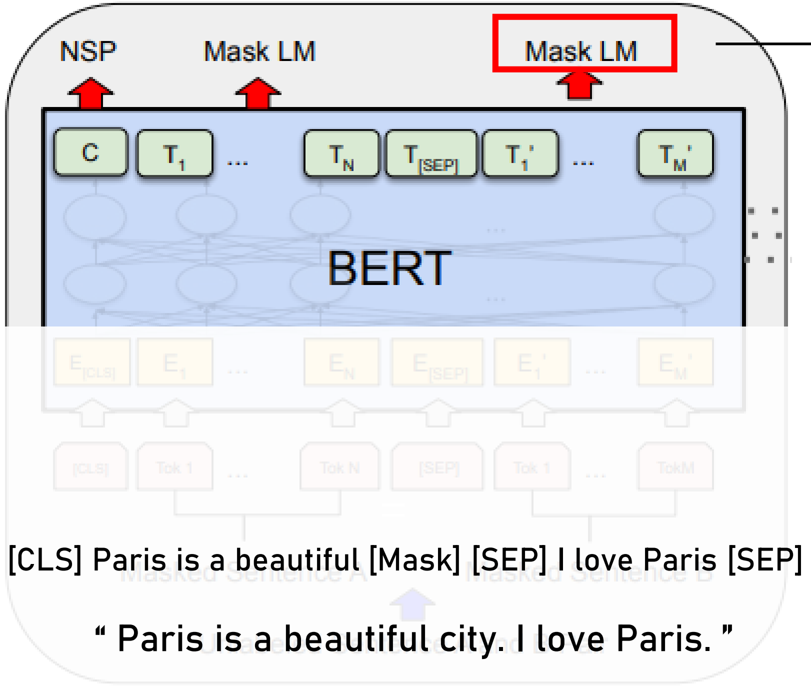


Pre-training

- **Pre-training, Fine-tuning mismatch**
- **Mask ratio**
  - 80% of the time: my dog is hairy → my dog is **[MASK]**
  - 10% of the time: my dog is hairy → my dog is **apple**
  - 10% of the time: my dog is **hairy**

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

BERT: MLM(Masked Language Model)

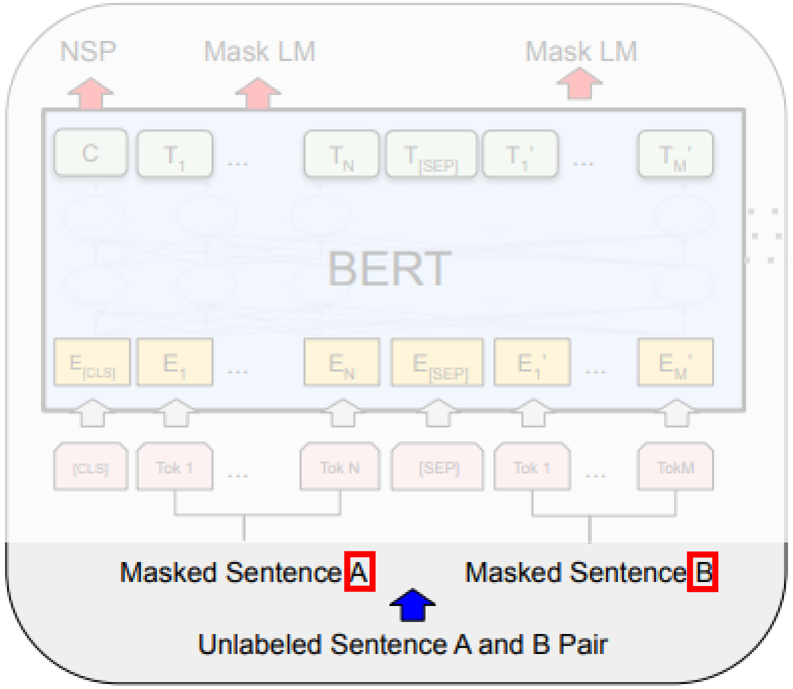


Pre-training

beautiful	0.0
…	
city	0.9
paris	0.0
love	0.1
and	0.0

SoftMax

BERT: NSP(Next Sentence Prediction)



Pre-training

- **Many important downstream tasks such as Question Answering (QA) and Natural Language Inference (NLI) are based on understanding the relationship between two sentences.**
- **50%: IsNext 50%: NotNext**

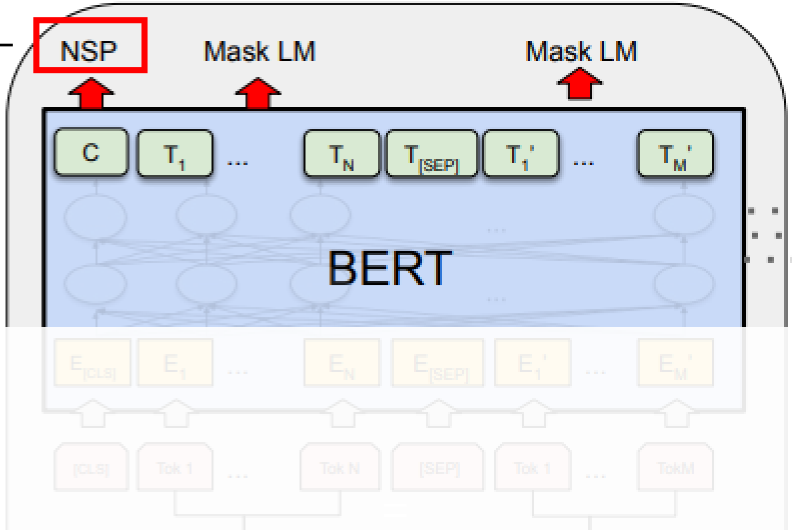
Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]
Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]
Label = NotNext

BERT: NSP(Next Sentence Prediction)

isNext	0.9
NotNext	0.1

SoftMax

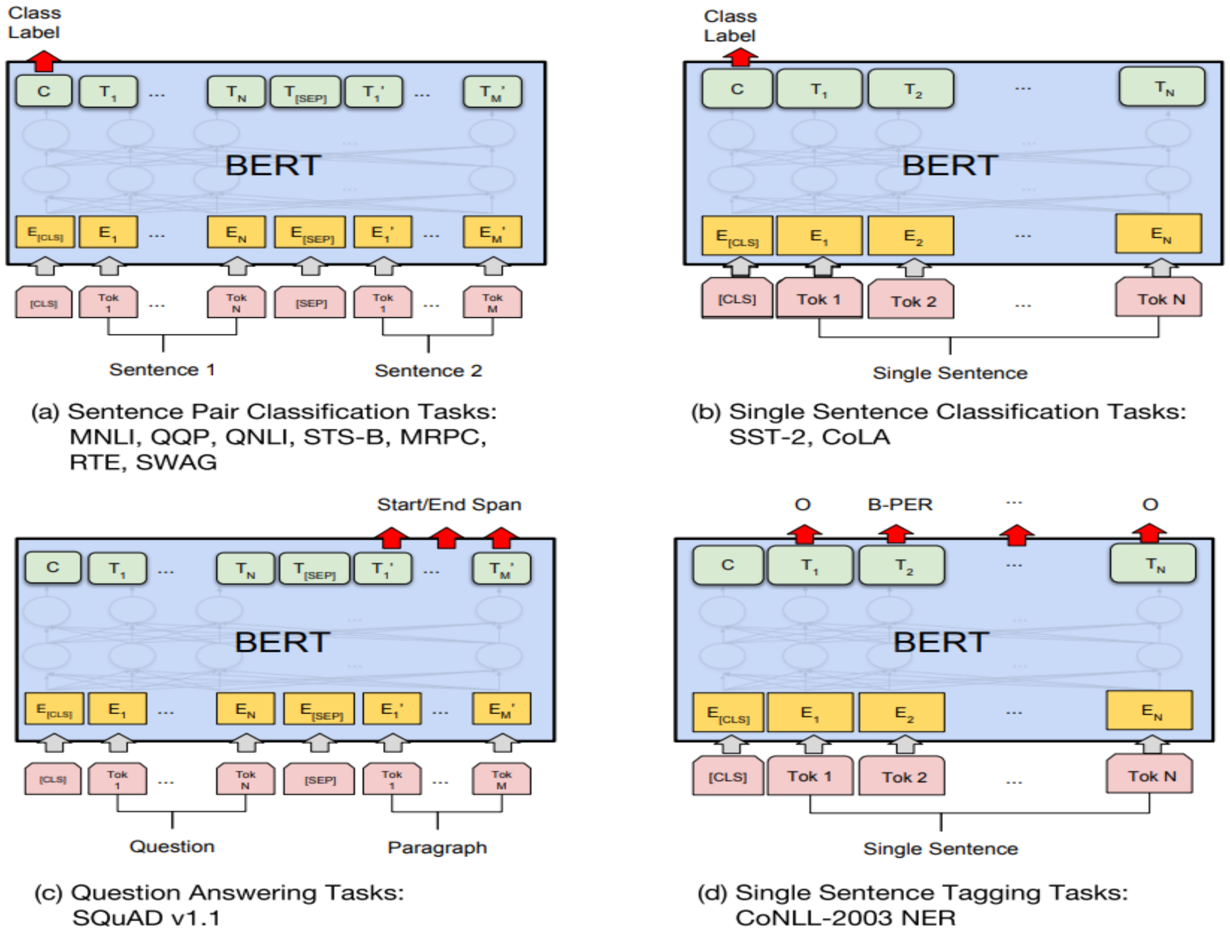


[CLS] Paris is a beautiful [Mask] [SEP] I love Paris [SEP]

“ Paris is a beautiful city. I love Paris. ”

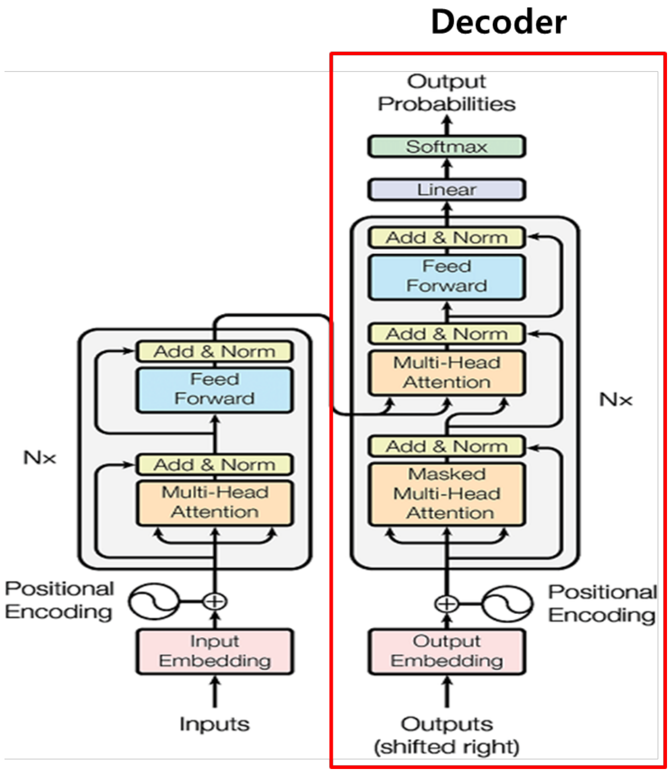
Pre-training

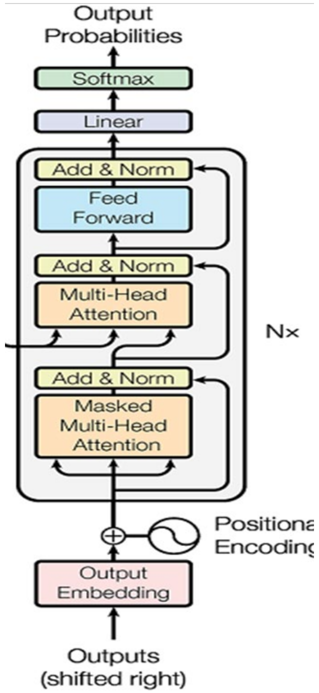
Finetuning

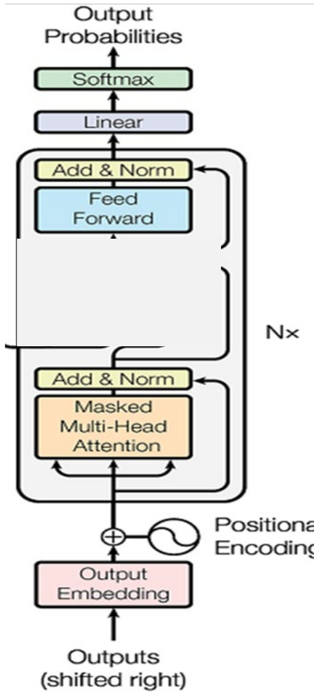


We can fine-tune by just adding output layer

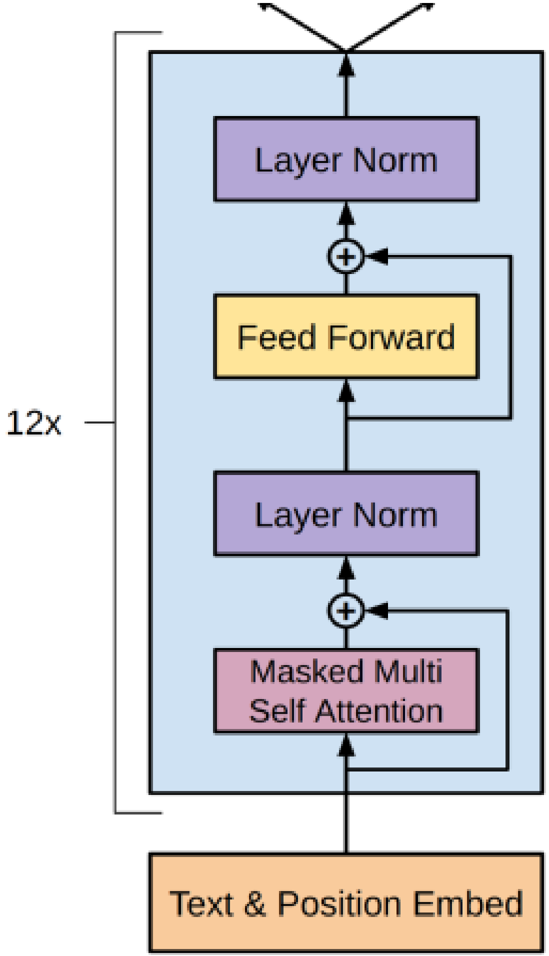
GPT-1


Decoder
Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Input
Embedding
Inputs
Positional
Encoding
Output
Embedding
Outputs
(shifted right)


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Output
Embedding
Outputs
(shifted right)

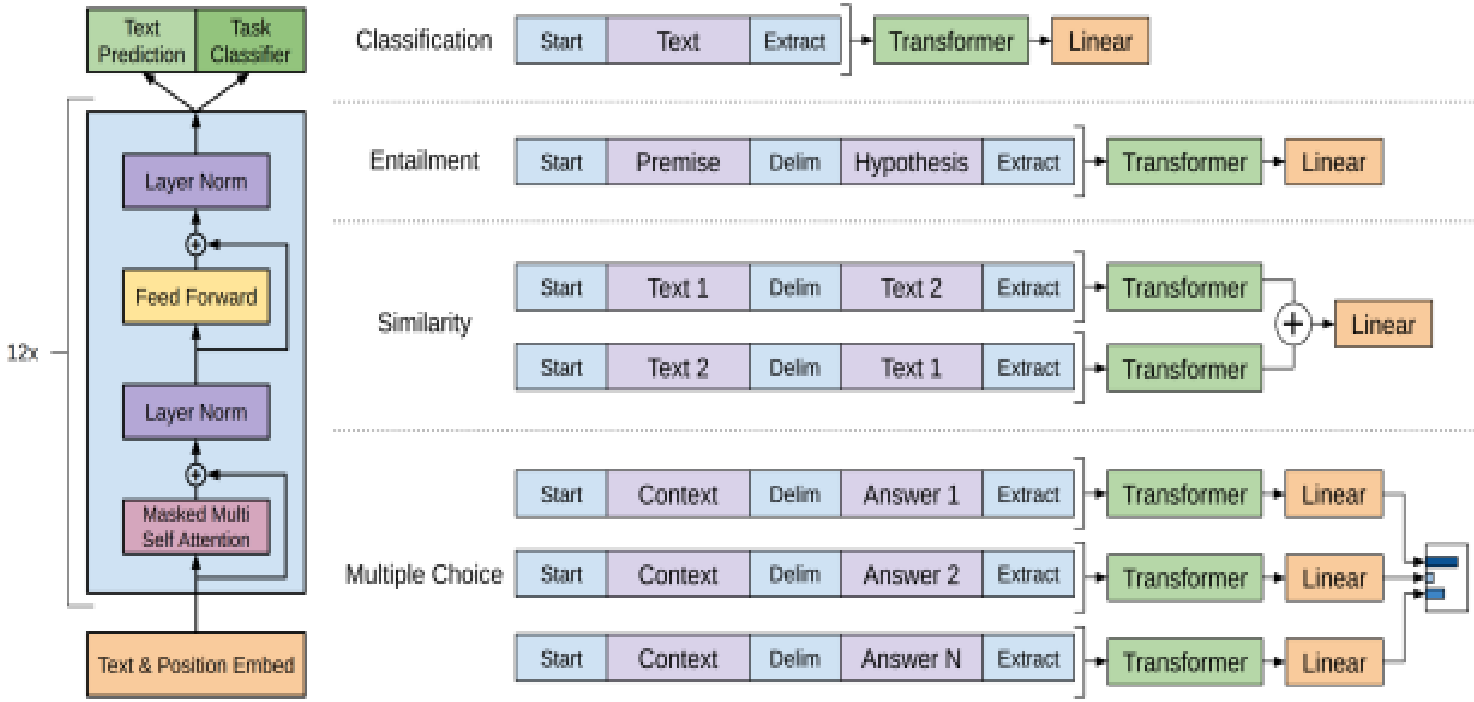

Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Masked
Multi-Head
Attention
Nx
Output
Embedding
Outputs
(shifted right)

GPT-1: Unsupervised pre-training



- Unsupervised corpus of tokens $U = \{u_1, \ldots, u_n\}$
- $L_1(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1} ; \theta)$
- $h_0 = UW_e + W_p$
- $h_l = transformer\_block(h_{l-1}) \; \forall i \in [1, n]$
- $P(u) = softmax(h_n W_e^T)$

GPT-1: Supervised fine-tuning



- Input tokens $x^1, \ldots, x^m$ and label y
- the final transformer block's activation h_l^m, which is then fed into an added linear output layer with parameters W_y to predict y:

$$P(y|x^1, \ldots, x^m) = softmax\left(h_l^m W_y\right)$$

$$L_2(C) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)$$

GPT-1, 2, 3

- **GPT-1**: Improving Language Understanding by Generative Pre-Training

  generative pre-training, **auxiliary fine-tuning**

- **GPT-2**: Language Models are Unsupervised Multitask Learners

  bigger model, **zero-shot learning**

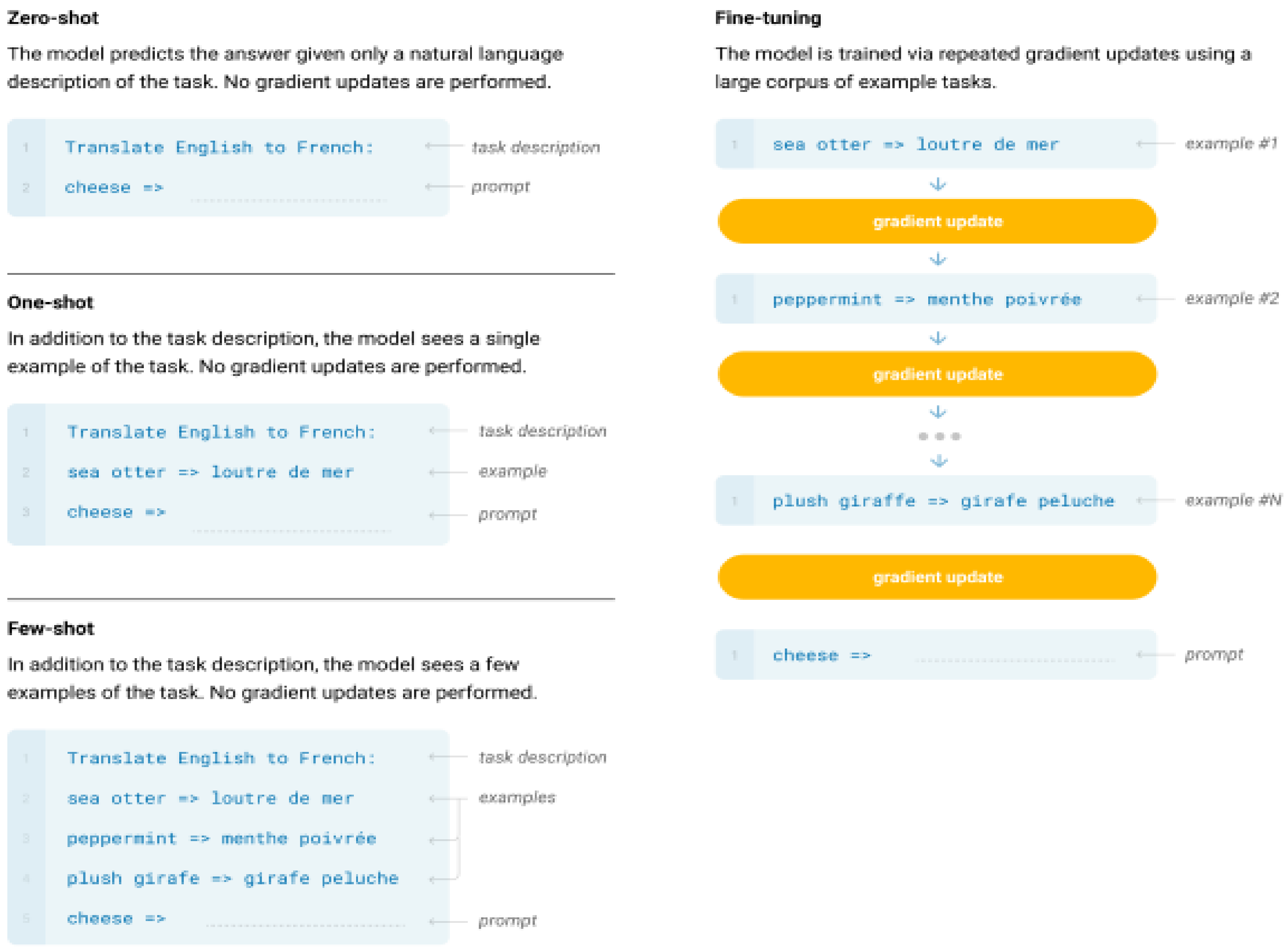
- **GPT-3**: Language Models are Few-Shot Learners

  biggest model, **few-shot learning**

GPT-1, 2, 3

	GPT-1	GPT-2	GPT-3
Parameters	117 Million	1.5 Billion	175 Billion
Decoder Layers	12	48	96
Context Token Size	512	1024	2048
Hidden Layer	768	1600	12288
Batch Size	64	512	3.2M

GPT-3: Downstream Task



Code

Start token ? End token ?

BERT

<CLS> context <SEP> question <SEP>

SQuAD2.0


```
{
    "qas": [
        {
            "question": "What is the law named that defines a charge moving through a magnetic field?",
            "id": "57378c9b1c456719005744a7",
            "answers": [
                {
                    "text": "Lorentz's Law",
                    "answer_start": 139
                },
                {
                    "text": "Lorentz's Law",
                    "answer_start": 139
                },
                {
                    "text": "Lorentz's Law",
                    "answer_start": 139
                },
                {
                    "text": "Lorentz's Law",
                    "answer_start": 139
                }
            ],
            "is_impossible": false
        },
        {
```


context: Through combining the definition of electric current as the time rate of change of electric charge, a rule of vector multiplication called **Lorentz's Law** describes the force on a charge moving in a magnetic field. The connection between electricity and magnetism allows for the description of a unified electromagnetic force that acts on a charge. This force can be written as a sum of the electrostatic force (due to the electric field) and the magnetic force (due to the magnetic field). Fully stated, this is the law:

1. 정답이 없는 경우
2. context 길이와 answer 길이가 안 맞는 경우
3. 길이가 max len 보다 긴 경우

ChatGPT

https://openai.com/blog/chatgpt/