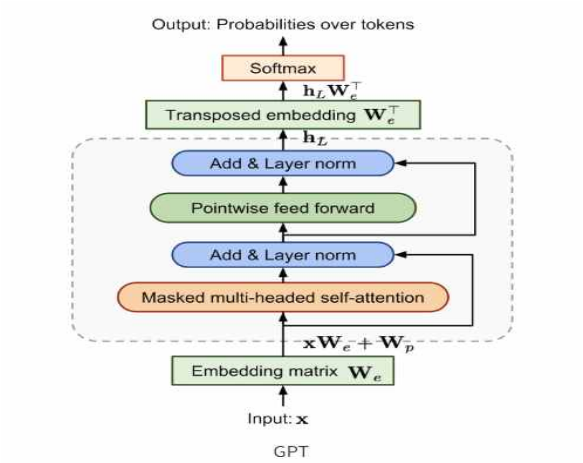
논문 분석 리포트

24기 정석호, 이건, 박동연, 이동진

[GPT-1 : Improving Language Understanding by Generative Pre-Training]

GPT-1의 가장 큰 구조적 특징은 전통적인 Transformer에서 Encoder를 제외하고 Decoder만으로 구성되었다는 것입니다. 기존의 Decoder는 크게 Self-Attention과 Cross-Attention으로 구성되는데, 이 중 Encoder와 Decoder의 인풋을 Cross Attention하는 Cross Self Attention부분을 제외한 부분만을 사용합니다. 즉 GPT-1은 Decoder의 Self-Attention 부분만으로 구성되어 간결한 모델구조를 가져 연산량이 크게 줄어듭니다. 이와 동시에 각 토큰이 이전의 모든 토큰과 어떻게 관련되는지를 파악하고, 이를 기반으로 다음 토큰을 예측하여 순차적으로 입력 텍스트를 처리하면서, 각 시점에서 전체 입력 시퀀스의 맥락을 고려할 수 있게 됩니다.



<GPT-1 모델 아키텍처>

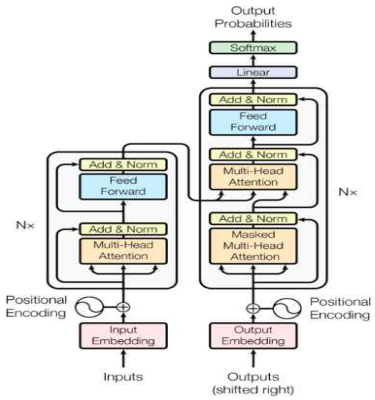


Figure 1: The Transformer - model architecture.

<기존 트랜스포머 아키텍처>

이처럼 GPT-1 모델은 트랜스포머 아키텍처의 효과적인 활용과 함께 자연어 처리 분야에서 많은 발전을 가져왔습니다. 구체적으로 두 문장의 유사한 정도를 파악하는 의미적 유사성 평가 작업과 주어진 텍스트의 범주를 분류하는 텍스트 분류 작업에서 좋은 성능을 보입니다. 뿐만 아니라 전제의 참, 거짓, 중립 여부를 판별하는 자연어 추론 작업과 질문 답변 작업에서도 뛰어난 성능을 보여 다양한 자연어 처리 작업 분야에 활용되어 왔습니다.

GPT-1은 기존 언어 모델에서 거의 활용하지 못하던 Unlabeled data를 Generative Pre-training 방식을 사용하여 보다 강력한 Pre-Training을 진행합니다. Pre-training과 Fine Tuning을 모두 거친 GPT-1는 대부분의 데이터셋의 대부분의 태스크에서 뛰어난 성능을 보였습니다. 이렇듯 GPT-1 모델은 강력한 Pre-training 방법을 제안해 이후의 언어 모델에 새로운 패러다임을 제시했다는 의의가 있습니다. 또한 Fine tuning 과정이 Task별로 모델을 크게 변경하지 않고 진행할 수 있어 Pretraining의 효과를 더 강하게 누릴 수 있습니다. 하지만 이전의 다른 모델들에 비해 연산량이 너무 높고, Fine Tuning을 직접해주어야 하기 때문에 Task 하나하나 Fine tuning 해주기 어렵다는 한계점도 있습니다.

[BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding]
BERT의 아키텍쳐는 multi-layer bidirectional Transformer 인코더로 되어있습니다. 즉, 여러층의 양방향 transformer 인코더로 구성이 되어있습니다. 또한, BERTbase는 12개의 인코더, 768차원의 임베딩 벡터, 12개의 self-attention으로 구성이 되어있습니다. Large 모델의 경우 24개, 1024차원, 16개로 더욱 파라미터의 개수가 많아졌습니다. 또한, BERTbase의 경우 이후에 언급할 GPT와의 태스크 테스트를 위해 같은 개수의 파라미터로 만들어졌습니다.

BERT의 경우에는 MLM과 NSP두 가지를 위하여 transformer 인코더만을 사용하였는데, input을 포함한 그 구조는 다음과 같습니다. input에서는 wordpiece tokenizer를 이용하여 자주 등장하는 단어들은 그대로 단어집합에 사용하고, 자주 등장하지 않으면 더 작은 단위로 분리하여 단어집합에 추가합니다. 또한 각 단어의 위치정보를 알기 위해 positional 임베딩을 진행하며, 문장을 구분하기 위해 Segment 임베딩을 이용하였습니다. 즉, input 임베딩은 위 세 임베딩 값이 합산됩니다. 이후 인코더에서는 Multi-head Self-Attention, Add & Norm, FFNN, Add & Norm 총 네 종류의 층으로 이루어집니다. 첫 번째의 경우 여러 개의 self-Attention이 병렬적으로 학습 되어있는 구조로 쉽게 역할만 말하자면 문장의 문맥을 파악하여 '그거'나 'it'같은 단어가 어떤 단어를 지칭하는지 판단을 하는 등 각 단어가 서로 어느 정도의 가중치가 있는지 판단합니다. 두 번째와 네 번째 층의 경우 Multi-head Self-Attention의 출력 값과 기존 값을 서로 더하여 기존의 input정보의 손실을 줄여줍니다. 마지막으로 세 번째 층은 정보가 한 방향으로 흐르는 특징을 이용하여 역전파를 통해 학습을 진행합니다.

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
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

논문의 실험 결과를 통해 미루어 보았을 때 OpenAI GPT나 ELMo와 비교한 결과에서 MNLI-m, QNLI, MRPC, SST-2, SQuAD 등의 벤치마크 실험에서 모두 더 뛰어난 성능을 보임을 확인할 수 있습니다. 구체적으로, MNLI는 가설과 전제 두 문장 사이의 관계를 이해하는지 실험하며, QNLI의 경우 질문과 질문에 대한 문장 사이의 관계를, MRPC는 두 문장이 서로 동일한 의미를 가지고 있는지, SST-2는 문장의 감성분류를, SQuAD는 문맥에서 질문에 대한 답을 잘 하는지 확인하는 태스크들입니다. 즉, 이러한 태스크들에 있어 GPT를 비롯한 SOTA 이상의 성능을 낸 것을 보아 위의 태스크에 주로 사용 가능하다고 볼 수 있습니다.
BERT는 자연어 처리 분야에서 큰 변화를 가져왔습니다. BERT의 성공은 깊은 양방향 문맥 표현이 NLP 작업에서 중요하다는 것을 보여줍니다. 또한, BERT는 전이 학습의 효과를 최대화하여 다양한 NLP 작업에서 SOTA 성능을 달성하였습니다. 이로 인해 BERT는 다양한 NLP 작업과 응용 분야에서 널리 사용되고 있습니다. 하지만, 기존 모델들에 비해 연산량이 많고, 대량의 학습 데이터가 필요하다는 한계점도 있습니다.