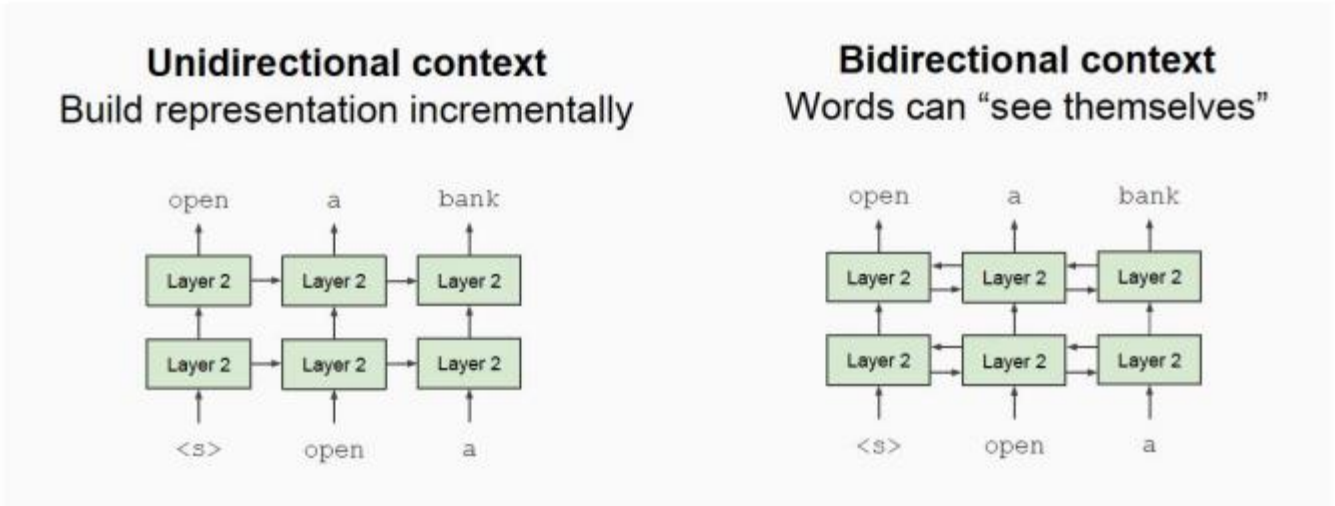
BERT: Pre-training of Deep Bidirectional Transformer for Language Understanding

- 2019 -

Concept

Pretrained Deep Bidirectional Representations



당연한 소리일 수 있지만 단어의 문맥적 의미를 표현하려면,

앞에 단어들만 보고 표현하는 것 보다는 뒤에 단어들도 함께 보면 좋겠지

Bidirectional Language Model

Forward LM

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}).$$

Backward LM

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N).$$

Maximize

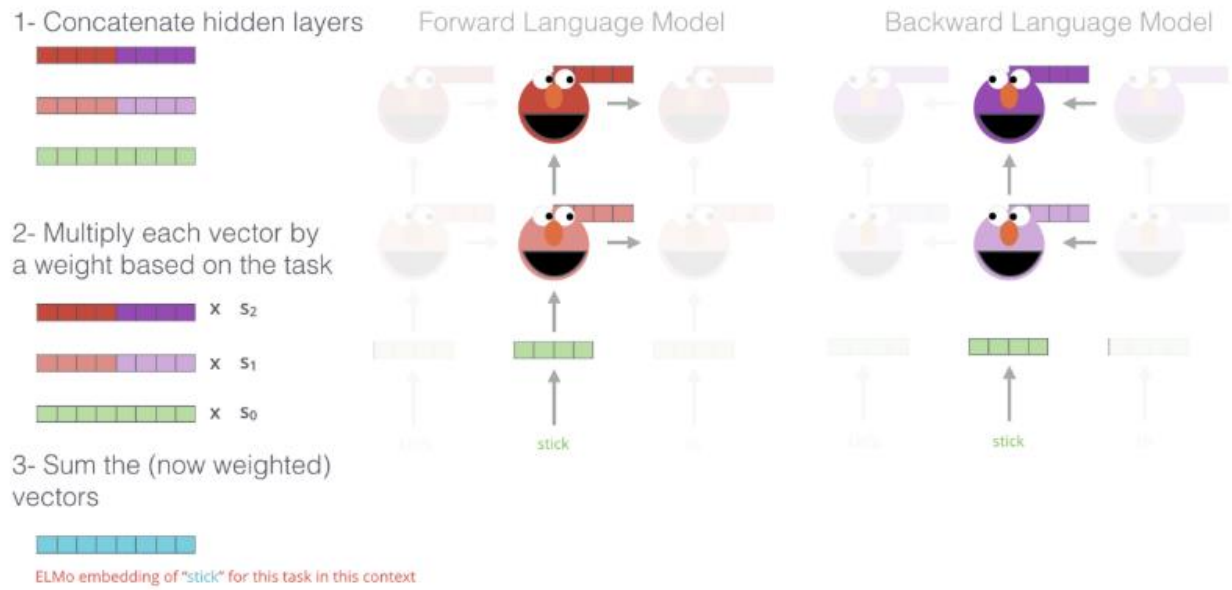
$$\sum_{k=1}^{N} (logp(t_k|t_1, \ldots, t_{k-1}; \theta) + logp(t_k|t_{k+1}, \ldots, t_N; \theta))$$

Previous work (ELMO)

ELMO Maximize

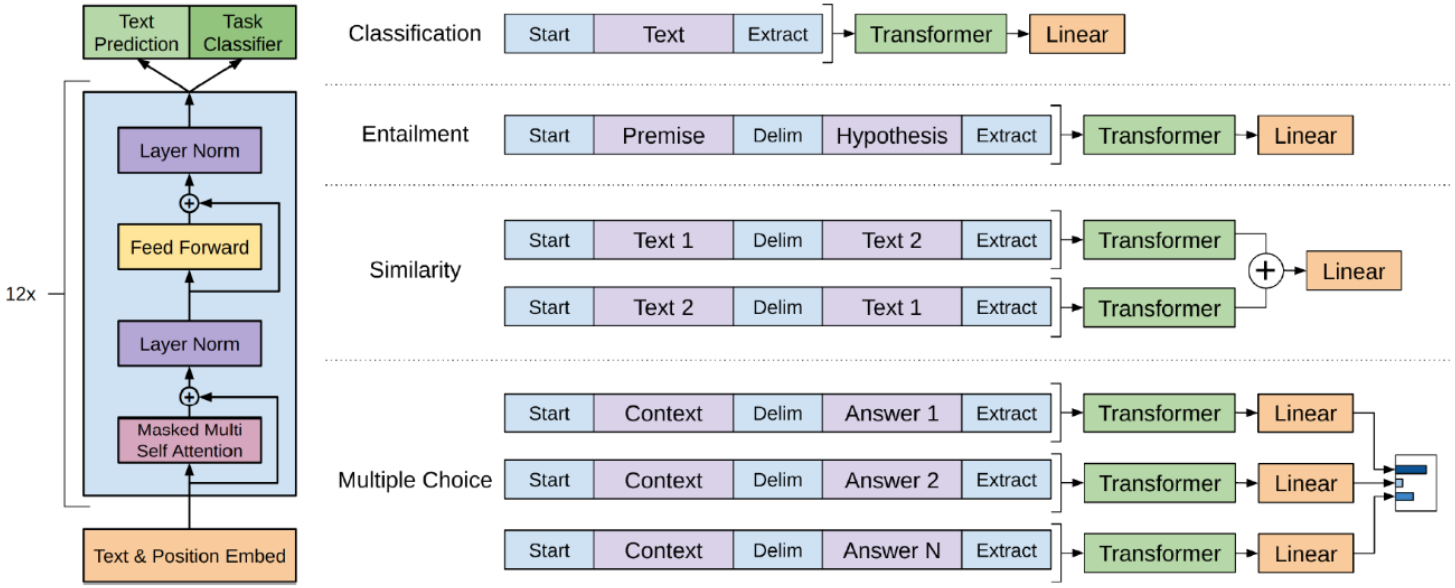
$$\sum_{k=1}^{N} \Big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big) .$$

ELMO가 새로운
word representation을
얻는 과정



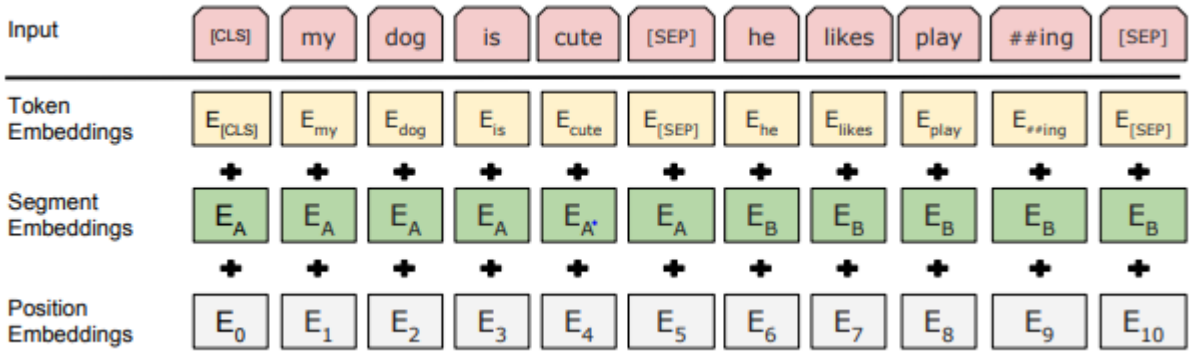
Previous work (GPT-1)

Improving Language Understanding by Generative Pre-Training



기본적으로는 Transformer Decoder 부분을 pre-training 하고
여러 task에서 파라미터를 미세 조정하여 응용한다.

Masked LM



[CLS] The first token of every sequence

- ✓ **마지막 layer에서 [CLS] 토큰의 값은 분류 모델의 input으로 사용**

[SEP] differentiate the sentences

- ✓ **마지막 layer에서 [CLS] 토큰의 값은 분류 모델의 input으로 사용**

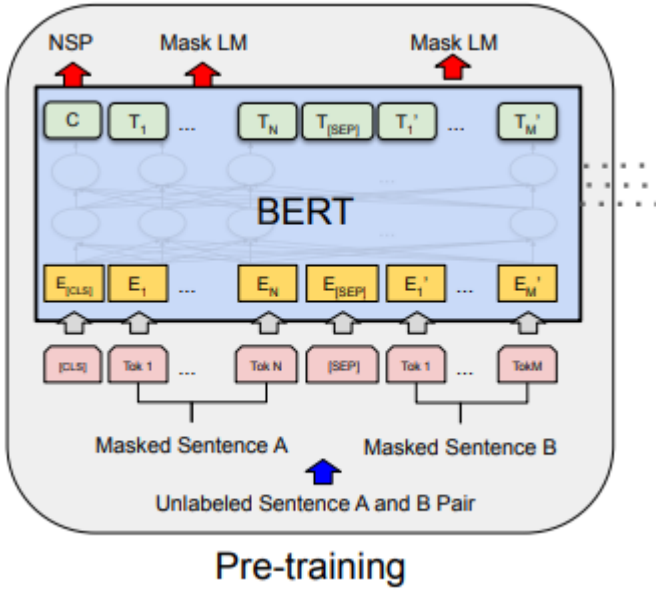
✓ 전체 Sequence 15%를 Masking 하여 predict 하는 Task

✓ 모든 예측 대상을 [Mask] 토큰으로 input에 넣어주는 것이 아니라

80%는 [Mask], 10%는 랜덤 단어, 10%는 원래 단어를 그대로 넣어주는 전략을 취한다.

-> 모델이 [Mask] 토큰이 아니면 아예 예측을 안 하는 방향으로 적합하지 않도록

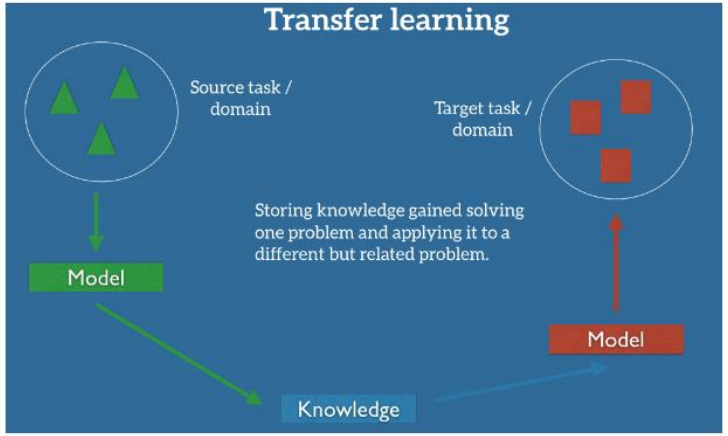
Next Sentence Prediction (NSP)



- ✓ 문장 간의 관계도 학습할 수 있도록 binarized next sentence prediction task 데이터를 사용하여 pre-train 진행 (이를 위해 Segment Embeddings 도입)
- ✓ 50%는 진짜로 이어지는 문장을 넣고, 50%는 랜덤 문장을 넣음
- ✓ [CLS] 토큰을 embedding 한 [C]에서 문장이 이어지는지 아닌지 예측

Applying pre-trained model to downstream tasks

Two types of transfer learning:



(1) Feature extraction

- ✓ **Only update the final layer weights from which we derive predictions**

(2) Fine – tuning:

- ✓ **Update all of the model's parameters for new task**