Unified Language Model Pre-training for Natural Language Understanding and Generation

Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (2019, Microsoft Research)

김유리

Index

Introduction

Introduction

- pre-training을 통해 다양한 NLP task가 발전됨
- Pre-trained LM은 context에 기반한 prediction word token들을 이용해 상황에 맞는 text를 representation
- 방대한 text data를 pre-training 한 후 downstream task에 맞게 모델을 fine-funning

Introduction

다양한 유형의 LM을 pre-train하기 위해 다양한 prediction task와 training objective들이 사용되고 있음

	ELMo	GPT	BERT	UNILM
Left-to-Right LM	✓	✓		✓
Right-to-Left LM	✓			✓
Bidirectional LM			✓	✓
Seq-to-Seq LM				✓

Table 1: Comparison between language model (LM) pre-training objectives. Seq-to-Seq is short for sequence-to-sequence.

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Introduction

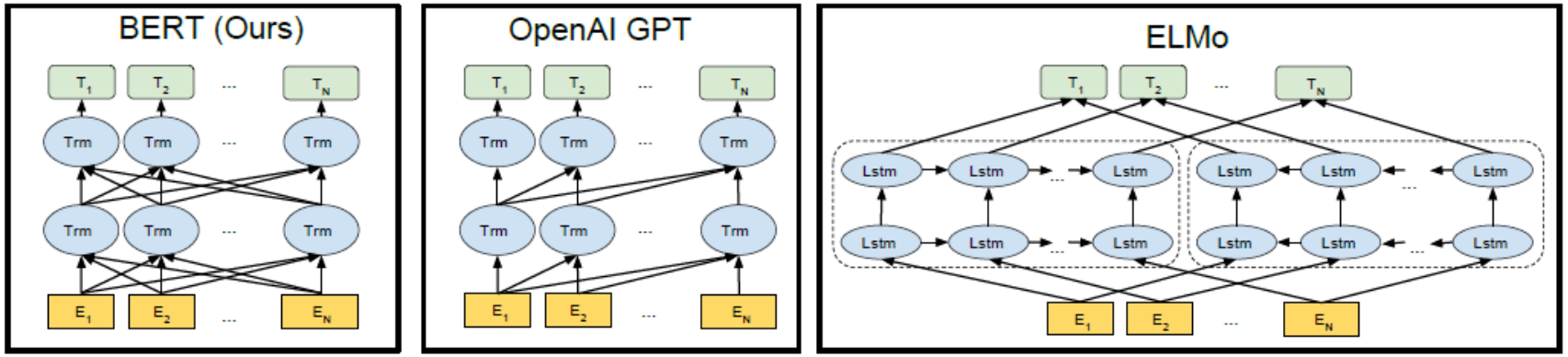


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

- **ELMo** : short-term memory networks 기반의 두개의 unidirectional LM을 학습
- **GPT** : left-to-right Transformer를 이용해 text sequence를 word단위로 predict
- **BERT** : bidirectional Transformer encoder를 사용해 masked word들을 predict하기 위해 오른쪽과 왼쪽 context를 fuse

"BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding", Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language, 2018)

Introduction

BERT

- 한 쌍의 텍스트의 관계를 명시적으로 모델링 할 수 있어 자연어 추론(Natural Language Inference)같은 pair-wise NLU task에 유리
- 광범위한 NLP task에서 성능을 향상시켰지만, **bidirectionality때문에 NLG(generation task)에 적용하기는 어려움**

→ 본 논문에서는 **NLU와 NLG 모두 적용** 가능한 **UNILM**(UNIfied pretrained Language Model)을 제안

Introduction

UNILM = Deep Transformer Network

대량의 text에 대해 3가지 유형의 unsupervised language modeling objectives로 활용됨(아래 표 참조)

Backbone Network	LM Objectives of Unified Pre-training	What Unified LM Learns	Example Downstream Tasks
Transformer with shared parameters for all LM objectives	Bidirectional LM	Bidirectional encoding	GLUE benchmark Extractive question answering
	Unidirectional LM	Unidirectional decoding	Long text generation
	Sequence-to-Sequence LM	Unidirectional decoding conditioned on bidirectional encoding	Abstractive summarization Question generation Generative question answering

Table 2: The unified LM is jointly pre-trained by multiple language modeling objectives, sharing the same parameters. We fine-tune and evaluate the pre-trained unified LM on various datasets, including both language understanding and generation tasks.

LM을 위한 cloze tasks를 설계함 (mask된 단어는 context기반으로 predict됨)

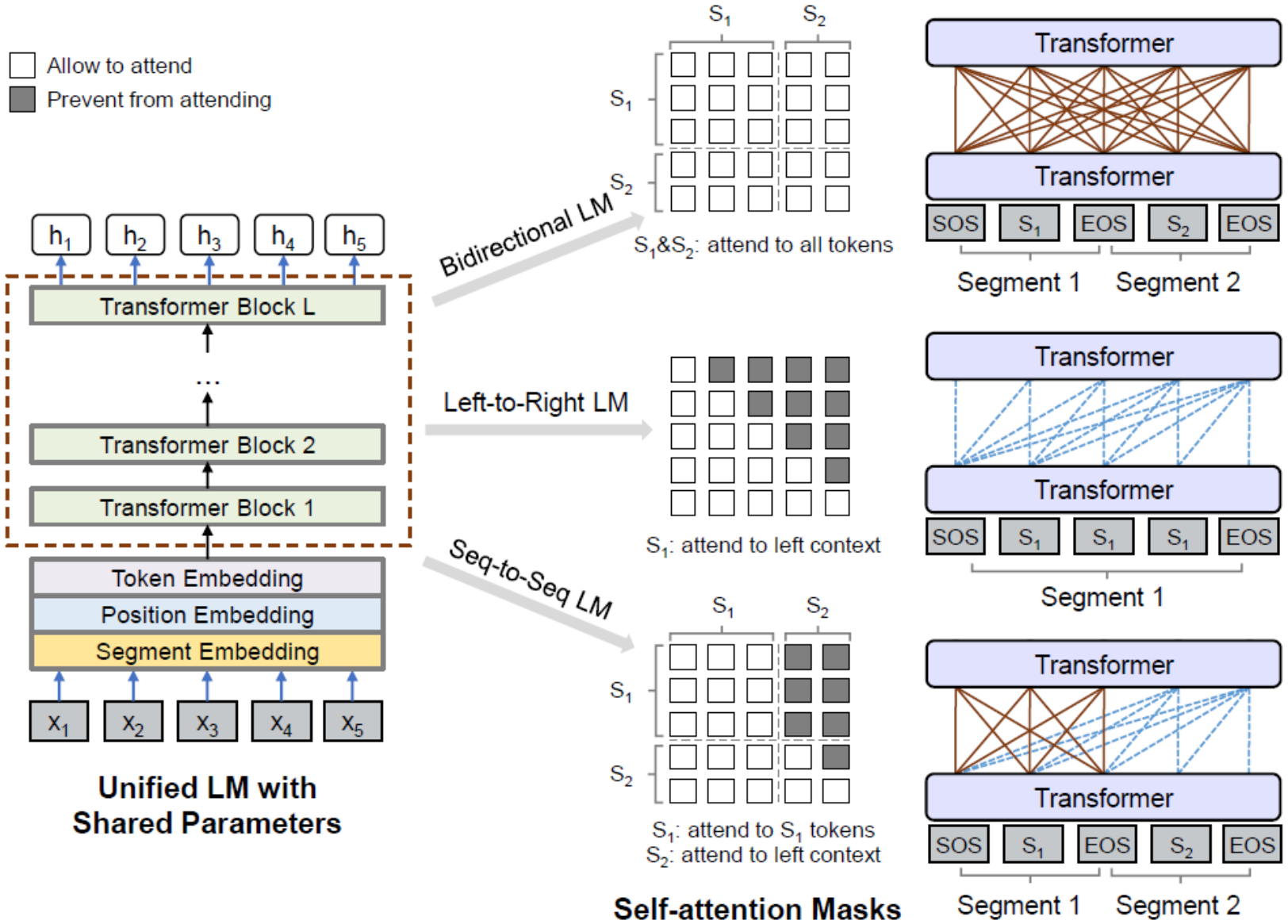
predict할 masked 단어의 context는?

- **Bidirectional LM** : 왼쪽 → 오른쪽 : 왼쪽 모든 단어로 / 오른쪽 → 왼쪽 : 오른쪽 모든 단어로 구성
- **Unidirectional LM** : 왼쪽 오른쪽 모든 단어로 구성
- **seq2seq LM** : predict 할 target sequence 의 context 는 source sequence의 모든 단어들 + target sequence 의 왼쪽에 있는 단어들로 구성

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

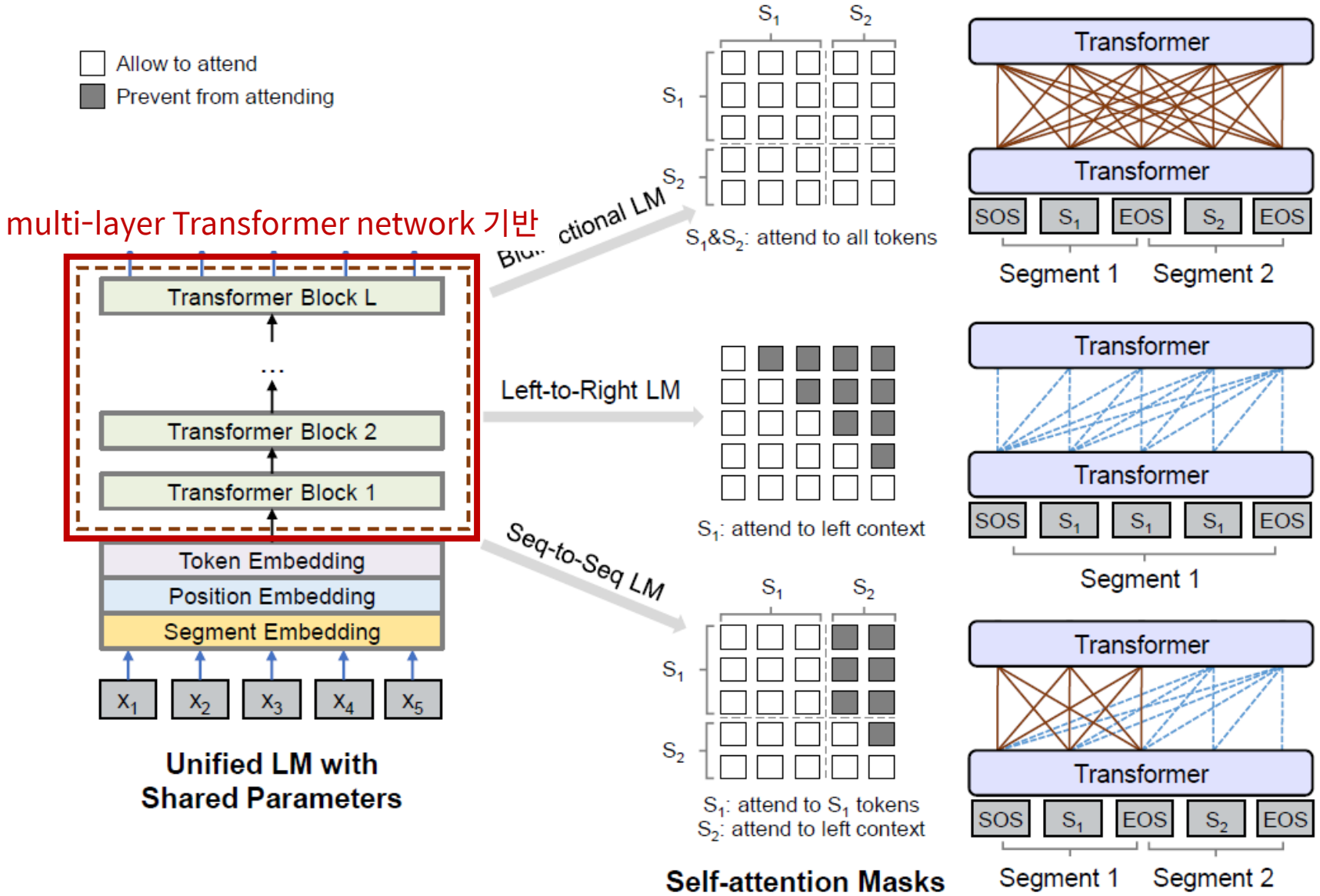
Methods

Methods
Architecture



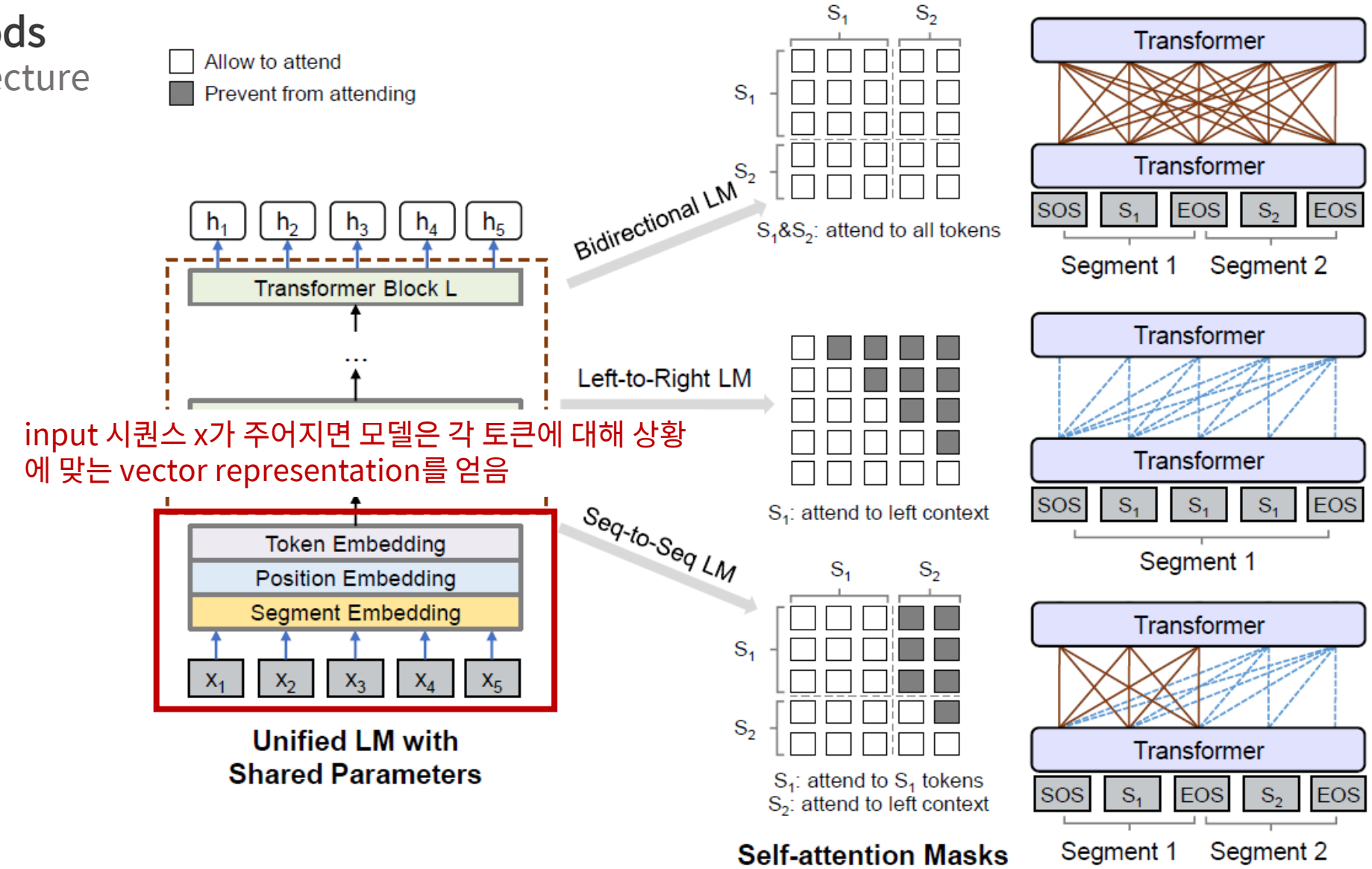
"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Methods
Architecture



"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

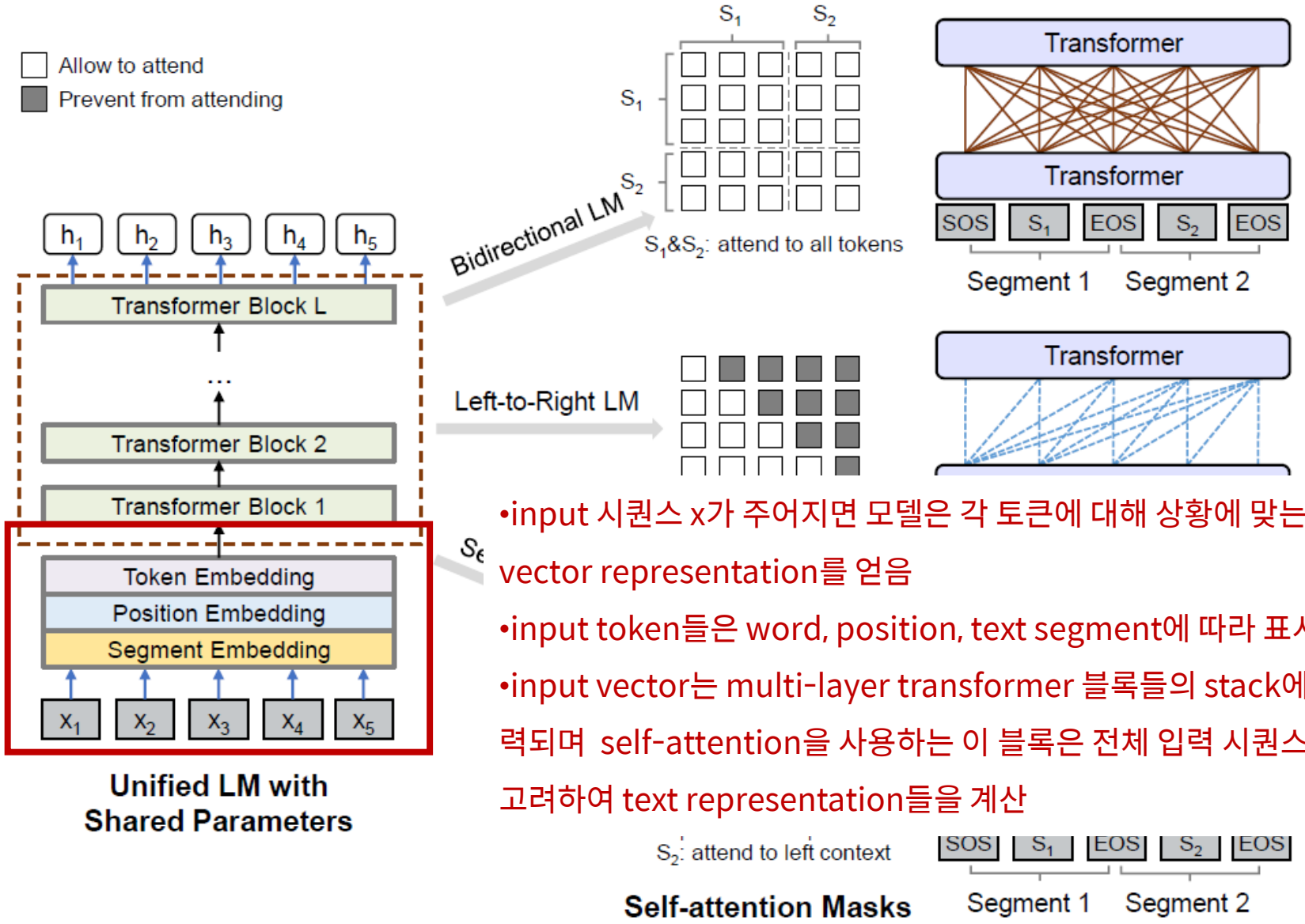
Methods
Architecture



input 시퀀스 x가 주어지면 모델은 각 토큰에 대해 상황에 맞는 vector representation를 얻음

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

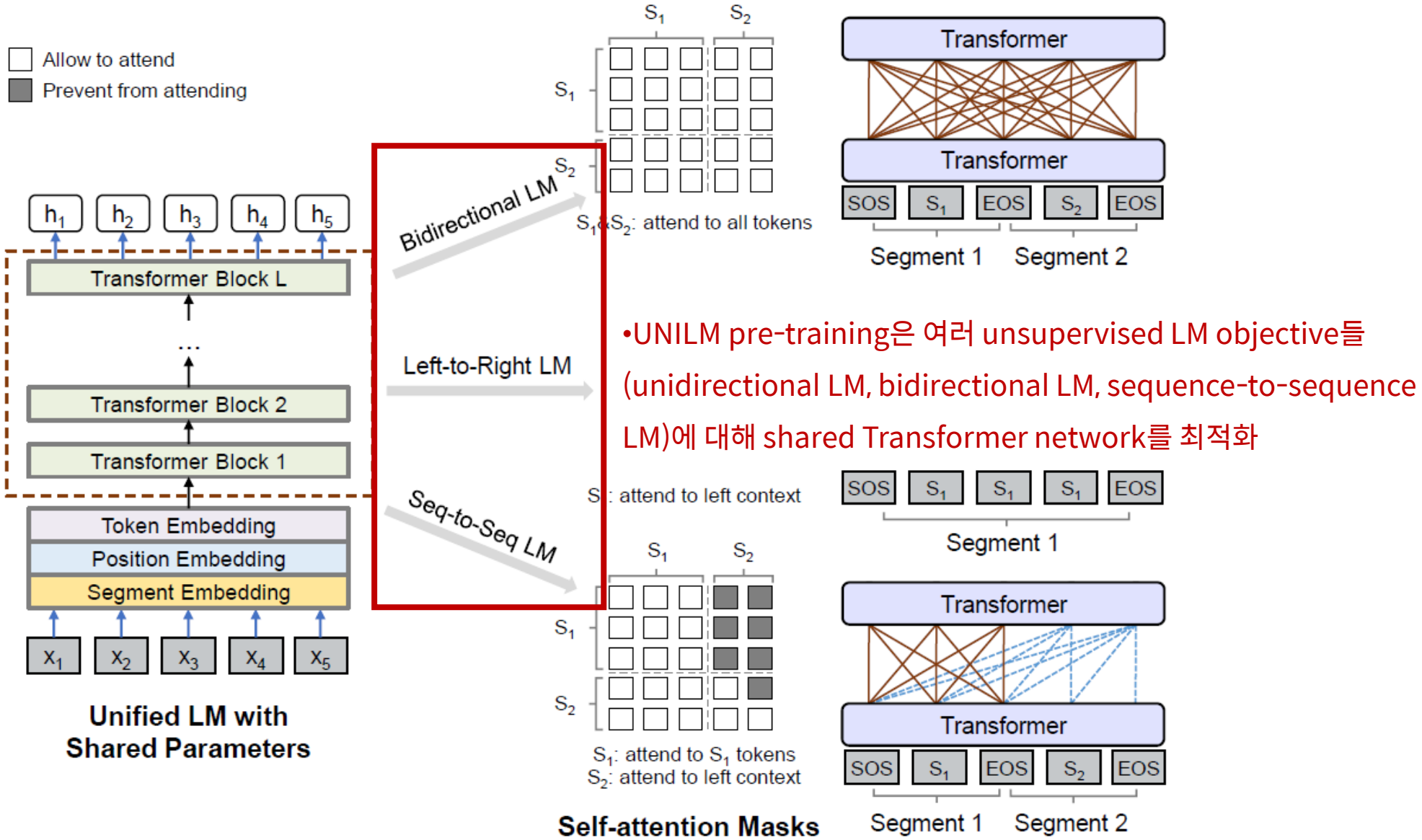
Methods
Architecture



- input 시퀀스 x가 주어지면 모델은 각 토큰에 대해 상황에 맞는 vector representation를 얻음
- input token들은 word, position, text segment에 따라 표시됨
- input vector는 multi-layer transformer 블록들의 stack에 입력되며 self-attention을 사용하는 이 블록은 전체 입력 시퀀스를 고려하여 text representation들을 계산

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Methods
Architecture



•UNILM pre-training은 여러 unsupervised LM objective들 (unidirectional LM, bidirectional LM, sequence-to-sequence LM)에 대해 shared Transformer network를 최적화

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

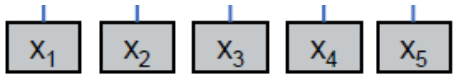
Methods
Architecture

Allow to attend
Prevent from attending

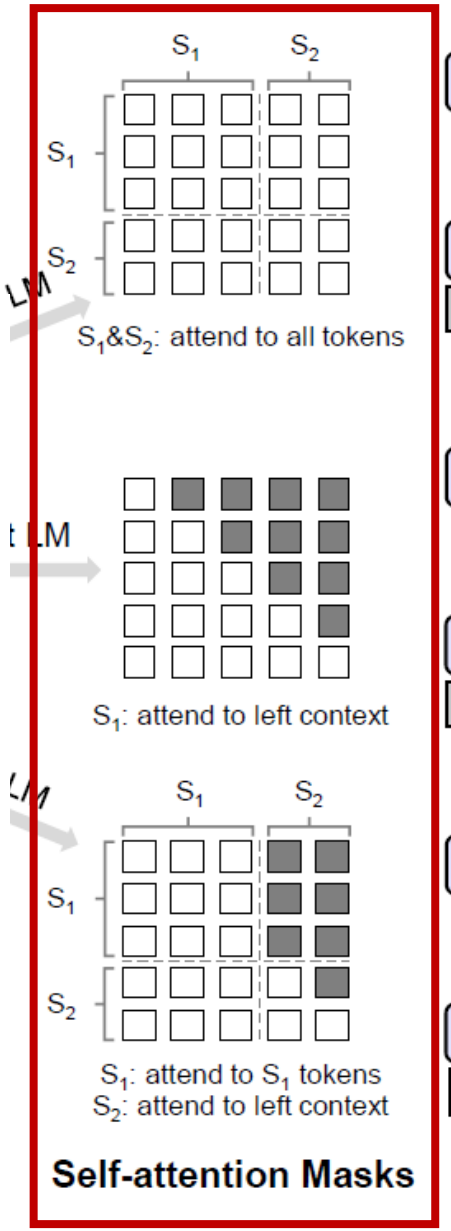
•predict될 word token의 context에 대한 접근을 제어하기 위해 self-attention에 다른 mask들을 사용함

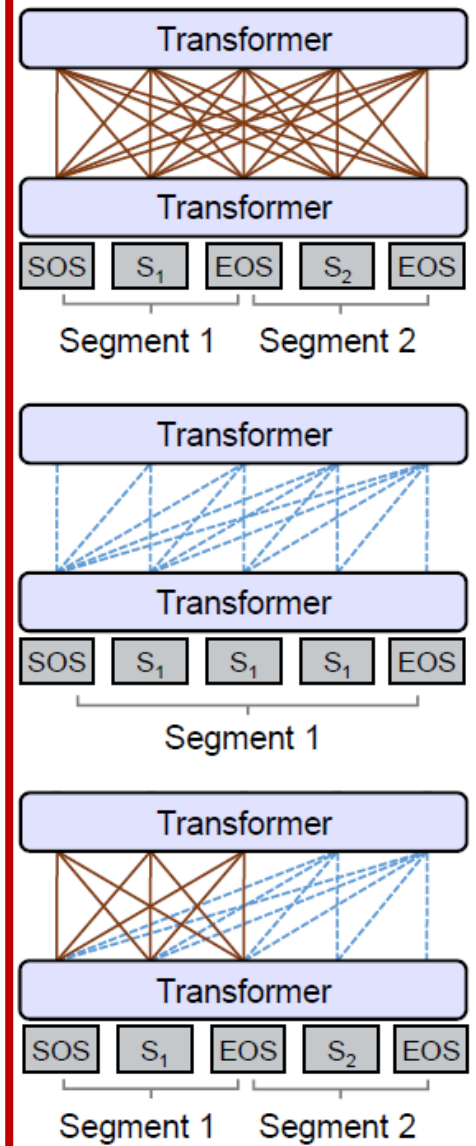
•다시 말해, masking을 사용해 contextualized representation을 계산할 때 토큰이 얼마나 많은 context에 영향을 주는지 제어

→ Unified LM이 pre-train되면 다양한 downstream task들에 대한 task-specific data를 fine-tune할 수 있음



Unified LM with Shared Parameters

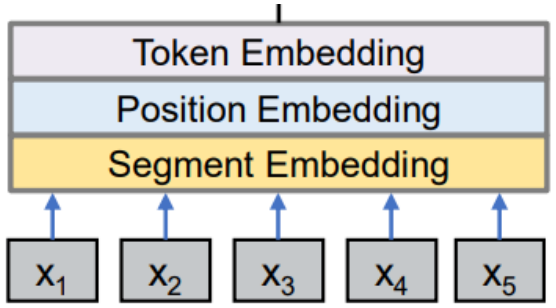




"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Methods

Input Representation



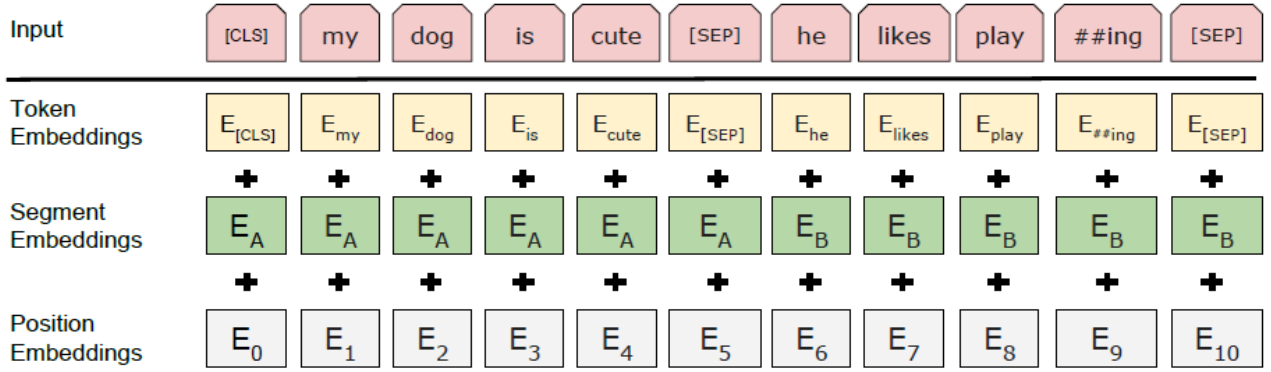


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

- input x는 word sequence (unidirectional LM : text segment / bidirectional LM, seq2seq LM :한 쌍의 segment)
- input을 시작할 때 [SOS] 토큰(special start-of-sequence) 추가 / 각 segment 끝에 [EOS]토큰(special end-of-sequence)추가
- 토큰이 한 쌍의 segments의 경계를 표시 함
- corresponding output vector를 전체 출력으로 사용
- [EOS]토큰은 NLU task에서 문장 간의 경계를 표시하면서, NLG task를 위한 모델 학습 시 decoding process를 종료하는데 사용됨
- input representation은 BERT와 같음 ("forecasted" → "forecast", "##ed" ##는 하나의 단어에 포함됨을 나타냄)
- 각 input token의 vector representation은 corresponding token embedding + position embedding + segment embedding (sum)
- UNILM은 여러가지 LM task들을 사용하여 학습되기 때문에 segment embedding이 어떤 LM 인지 구분하는 식별자 역할을 함 (각 LM마다 다른 segment embedding을 사용하기 때문)

Methods
Backbone Network : Multi-Layer Transformer

$$\mathbf{H}^l = \text{Transformer}_l(\mathbf{H}^{l-1}), l \in [1, L]$$

$$\{\mathbf{x}_i\}_{i=1}^{|x|} \longrightarrow \mathbf{H}^0 = [\mathbf{x}_1, \cdots, \mathbf{x}_{|x|}] \longrightarrow \mathbf{H}^l = [\mathbf{h}_1^l, \cdots, \mathbf{h}_{|x|}^l]$$

Input vector

압축

L-layer Transformer 사용
contextual representation으로 인코딩

- 각 Transformer block안에는 여러 개의 self-attention head가 존재
- 그 head들이 이전 layer의 output 벡터를 aggregate

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Methods
Backbone Network : Multi-Layer Transformer

l번째 Transformer layer에서 self-attention head 의 출력은 아래 식을 통해 계산

$$\mathbf{Q} = \mathbf{H}^{l-1}\mathbf{W}_l^Q, \quad \mathbf{K} = \mathbf{H}^{l-1}\mathbf{W}_l^K, \quad \mathbf{V} = \mathbf{H}^{l-1}\mathbf{W}_l^V \tag{1}$$

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases} \tag{2}$$

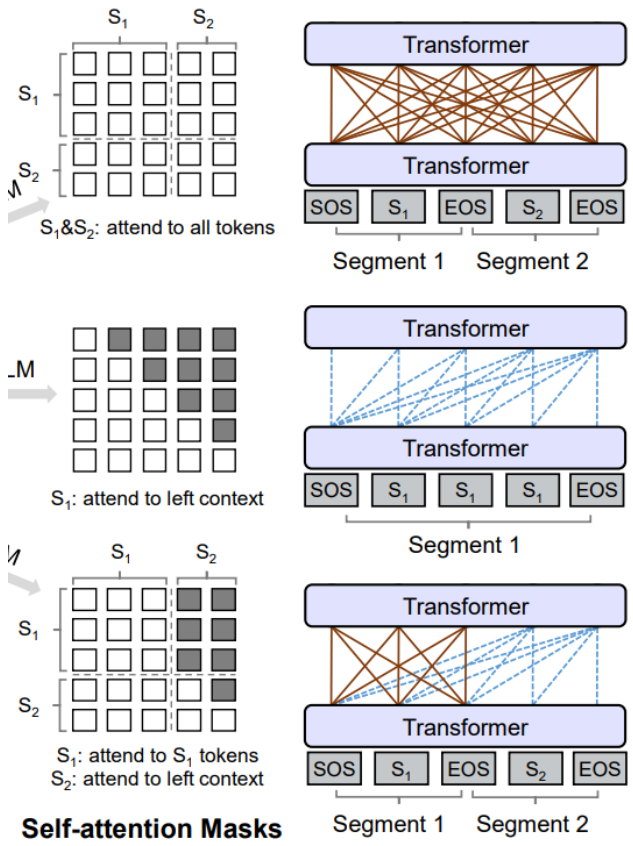
$$\mathbf{A}_l = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V}_l \tag{3}$$

이전 layer의 output인 $\mathbf{H}^{l-1} \in \mathbb{R}^{|x| \times d_h}$ 는 각각 매개 변수 행렬 $\mathbf{W}_l^Q, \mathbf{W}_l^K, \mathbf{W}_l^V \in \mathbb{R}^{d_h \times d_k}$ 을 이용해
query, key, value들이 3배로 linearly project 되며, mask 행렬 $\mathbf{M} \in \mathbb{R}^{|x| \times |x|}$ 은 한 쌍의 token이 서로 attend할 수 있는지 여부 결정

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Methods

Backbone Network : Multi-Layer Transformer



- 그림처럼 contextualized representation을 계산할 때 token이 어떤 context에 영향을 줄지를 조절하기 위해 mask 행렬인 M 사용
- bidirectional LM을 예로 들면, mask 행렬의 요소는 모두 0이며 모든 token이 서로 access할 수 있음을 나타냄

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Methods
Pre-training Objectives

- 서로 다른 language modeling을 위해 설계된 네가지 cloze task를 사용해 UNILM을 pretain
- cloze task에서 임의의 일부 WordPiece token을 선택해 special token [MASK]로 바꿈
- Transformer network에서 계산된 output 벡터를 softmax classifier을 이용해 masked token을 predict함
- UNILM의 parameter들은 predicted token들과 original token들을 이용해 cross-entropy loss를 최소화하는 방향으로 학습 함

Methods
Pre-training Setup

- 1개의 training batch를 학습 시킬 때, 시간 기준으로 1/3은 bidirectional LM, 1/3은 Sseq2seq LM, 1/6은 left to right unidirectional LM, 1/6 right to left unidirectional LM을 사용함
- Model 아키텍처는 BERTlarge 사용
- gelu activation (like GPT)
- 24 layer transformer, 1024 hidden size, 16 self-attention head (340M parameter)

Methods
Fine-tuning on Downstream Tasks

- NLU : BERT를 레퍼런스로 함 / [EOS]토큰은 문장 간의 경계를 표시
- NLG : S2 문장의 token만 마스킹 / [EOS]토큰은 모델 학습 시 decoding process를 종료하는데 사용됨

Experiments

Experiments

NLU, NLG task에 대해 실험 진행

- **NLU** : GLUE benchmark, extractive question answering
- **NLG** : abstractive summarization, question generation, generative question answering, dialog response generation

Experiments
Abstractive Summarization

	RG-1	RG-2	RG-L
Extractive Summarization			
LEAD-3	40.42	17.62	36.67
Best Extractive [27]	43.25	**20.24**	39.63
Abstractive Summarization			
PGNet [37]	39.53	17.28	37.98
Bottom-Up [16]	41.22	18.68	38.34
S2S-ELMo [13]	41.56	18.94	38.47
UNILM	**43.33**	20.21	**40.51**

Table 3: Evaluation results on CNN/DailyMail summarization. Models in the first block are extractive systems listed here for reference, while the others are abstractive models. The results of the best reported extractive model are taken from [27]. RG is short for ROUGE.

	RG-1	RG-2	RG-L
10K Training Examples			
Transformer [43]	10.97	2.23	10.42
MASS [39]	25.03	9.48	23.48
UNILM	**32.96**	**14.68**	**30.56**
Full Training Set			
OpenNMT [23]	36.73	17.86	33.68
Re3Sum [4]	37.04	19.03	34.46
MASS [39]	37.66	18.53	34.89
UNILM	**38.45**	**19.45**	**35.75**

Table 4: Results on Gigaword abstractive summarization. Models in the first block only use 10K examples for training, while the others use 3.8M examples. Results of OpenNMT and Transformer are taken from [4, 39]. RG is short for ROUGE.

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Experiments
Question Answering

	EM	F1
RMR+ELMo [20]	71.4	73.7
BERT$_{\text{LARGE}}$	78.9	81.8
UNILM	**80.5**	**83.4**

Table 5: Extractive QA results on the SQuAD development set.

	F1
DrQA+ELMo [35]	67.2
BERT$_{\text{LARGE}}$	82.7
UNILM	**84.9**

Table 6: Extractive QA results on the CoQA development set.

	F1
Seq2Seq [35]	27.5
PGNet [35]	45.4
UNILM	**82.5**

Table 7: Generative QA results on the CoQA development set.

SQuAD, CoQA 모두 BERT 능가

EM : expectation maximization

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Experiments
Question Generation

	BLEU-4	MTR	RG-L
CorefNQG [11]	15.16	19.12	-
SemQG [50]	18.37	22.65	46.68
UNILM	**22.12**	**25.06**	**51.07**
MP-GSN [51]	16.38	20.25	44.48
SemQG [50]	20.76	24.20	48.91
UNILM	**23.75**	**25.61**	**52.04**

Table 8: Question generation results on SQuAD. MTR is short for METEOR, and RG for ROUGE. Results in the groups use different data splits.

	EM	F1
UNILM QA Model (Section 3.2)	80.5	83.4
+ UNILM Generated Questions	**84.7**	**87.6**

Table 9: Question generation based on UNILM improves question answering results on the SQuAD development set.

Experiments
GLUE Benchmark

Model	CoLA MCC	SST-2 Acc	MRPC F1	STS-B S Corr	QQP F1	MNLI-m/mm Acc	QNLI Acc	RTE Acc	WNLI Acc	AX Acc	Score
GPT	45.4	91.3	82.3	80.0	70.3	82.1/81.4	87.4	56.0	53.4	29.8	72.8
$\text{BERT}_{\text{LARGE}}$	60.5	**94.9**	89.3	86.5	**72.1**	86.7/**85.9**	**92.7**	70.1	65.1	39.6	80.5
UNILM	**61.1**	94.5	**90.0**	**87.7**	71.7	**87.0/85.9**	**92.7**	**70.9**	65.1	38.4	**80.8**

Table 11: GLUE test set results scored using the GLUE evaluation server.

"Unified Language Model Pre-training for Natural Language Understanding and Generation", Li Dong, Nan Yang, Wenhui Wang, Furu Wei et al. (NeurIPS 2019)

Conclusion

Conclusion

- parameter를 공유함으로써 여러가지 LM objectives에 대해 최적화 된 unified pre-training model을 제안 함
- bidirectional, unidirectional, sequence-to-sequence LMs들의 통합으로 NLU, generation task에 대해 모두 사용가능한 pre-tranined 된 LM을 fine-tune할 수 있음
- GLUE벤치마크와 두가지 question answering dataset 실험 결과에서 UNILM이 BERT보다 우세하게 나타남
- 또한, unified pre-trained LM은 세가지 자연어 generation task들에서 (CNN/DailyMail abstractive summarization, SQuAD question generation, CoQA generative question answering) 기존 SOTA모델보다 좋은 성능을 나타냄

Thank you