Latest NLP

강사 : 백병인

pi.paek@modulabs.co.kr

모두의연구소 Research Scientist

GPT-2가 뭐길래??

https://openai.com/blog/better-language-models/

https://openai.com/blog/better-language-models/

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns. While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic." Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

너무 글 잘쓰는 AI... 깜짝 놀란 개발자들, 비공개

조선일보 | 원우식 기자

입력 2019.02.20 03:01

소설·뉴스·학교 숙제 척척
논리적으로 문장 구성하고 새로운 문장 탁월하게 작성

미국의 비영리 인공지능(AI) 연구기관 '오픈 AI(Open AI)'가 새로 개발한 '글짓기 인공지능'의 글쓰기 실력이 너무 뛰어나 연구자들이 '악용이 우려된다'는 이유로 원천 기술을 비공개하기로 결정했다. 오픈 AI는 테슬라 CEO 일론 머스크 등 미국 IT 기업 대표들이 2015년 '인공지능을 통해 인류에

오늘의 등장 논문들

- CoVe : McCann et al., 2017
  - Learned in Translation: Contextualized Word Vectors
  - https://arxiv.org/abs/1708.00107
- ELMo : Peters et al., 2018
  - Deep contextualized word representations
  - https://arxiv.org/abs/1802.05365
- CVT : Clark et al., 2018
  - Semi-Supervised Sequence Modeling with Cross-View Training
  - https://arxiv.org/abs/1809.08370
- GPT : Radford et al., 2018
  - Improving Language Understanding by Generative Pre-Training
  - https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- BERT : Delvin et al., 2018
  - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  - https://arxiv.org/abs/1810.04805
- Transformer-XL : Dai et al., 2019
  - Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context
  - https://arxiv.org/abs/1901.02860
- GPT-2 : Radford et al., 2019
  - Language Models are Unsupervised Multitask Learners
  - https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf?forcedefault=true
- 그 외 참조된 논문들

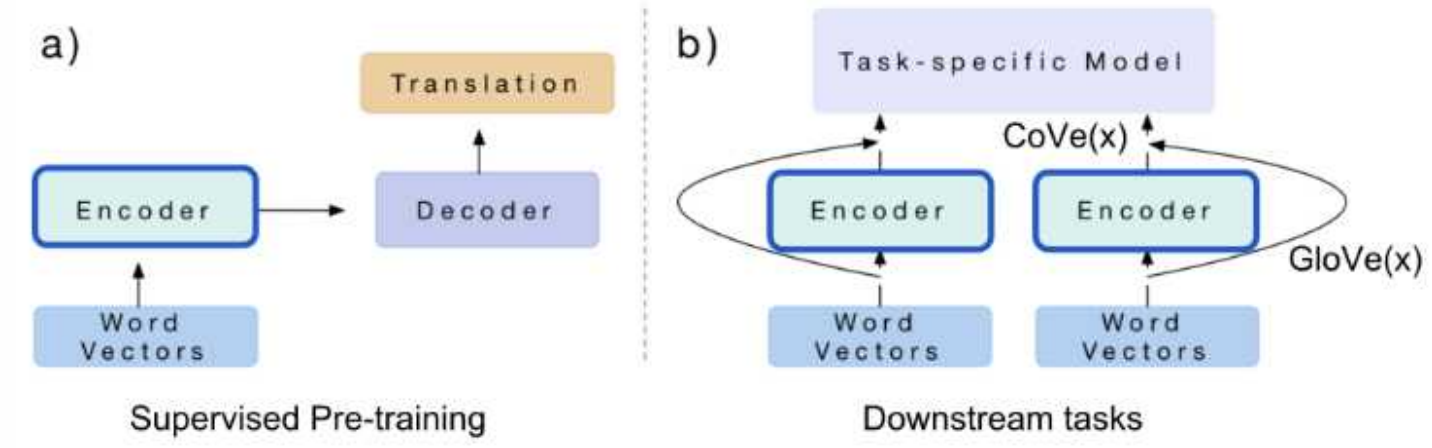
Summary

	Base model	pre-training	Downstream tasks	Downstream model	Fine-tuning
CoVe	seq2seq NMT model	supervised	feature-based	task-specific	/
ELMo	two-layer biLSTM	unsupervised	feature-based	task-specific	/
CVT	two-layer biLSTM	semi-supervised	model-based	task-specific / task-agnostic	/
ULMFiT	AWD-LSTM	unsupervised	model-based	task-agnostic	all layers; with various training tricks
GPT	Transformer decoder	unsupervised	model-based	task-agnostic	pre-trained layers + top task layer(s)
BERT	Transformer encoder	unsupervised	model-based	task-agnostic	pre-trained layers + top task layer(s)
GPT-2	Transformer decoder	unsupervised	model-based	task-agnostic	pre-trained layers + top task layer(s)

https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html

CoVe : 워드벡터에 맥락을 더하다

- Word Representation에 Context를 더하다.
- seq2seq 구조의 번역기에서, Encoder의 출력이 되는 context vector를 pretrain했다가 word vector와 결합해서 사용하면??



$$\mathrm{CoVe}(x) = \mathrm{biLSTM}(\mathrm{GloVe}(x))$$

$$v = [\mathrm{GloVe}(x); \mathrm{CoVe}(x)]$$

CoVe의 한계와 그 솔루션

- 한계 1 : Context Learning의 한계
  - context pretraining이 task-specific하다. (번역 코퍼스에 의존)
  - ELMo 와 CVT가 이 문제점을 개선한다.
- 한계 2 : Generality의 한계
  - Optimization이 task-specific하다. (Model이 task-specific)
  - 보다 general한 Language Model을 통해 다양한 문제를 한꺼번에 해결할 수는 없을까?
  - GPT, BERT가 이 문제점을 개선한다.

ELMo : Embeddings from Language Model

- Deep BiLSTM에 비지도학습으로 LM을 pretrain하면 그 안에서 state representation을 꺼내 활용할 수 있다.
- task별로 layer별 state의 가중치를 다르게 주도록 학습한다.
  - Semantic task는 Top Layer의 가중치가 높도록, Syntax task는 낮은 쪽 Layer의 가중치가 높도록 학습된다.

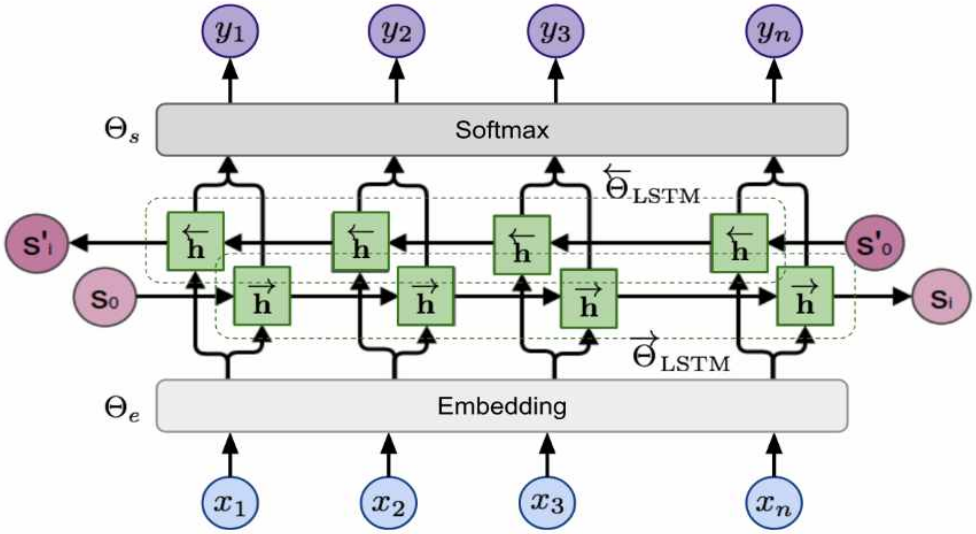


Fig. 3. The biLSTM base model of ELMo. (Image source: recreated based on the figure in "Neural Networks, Types, and Functional Programming" by Christopher Olah.)

$$\mathbf{h}_{i,\ell} = [\overrightarrow{\mathbf{h}}_{i,\ell}; \overleftarrow{\mathbf{h}}_{i,\ell}]$$

$$R_i = \{\mathbf{h}_{i,\ell} \mid \ell = 0, \ldots, L\}$$

$$v_i = f(R_i; \Theta^{\text{task}}) = \gamma^{\text{task}} \sum_{\ell=0}^{L} s_i^{\text{task}} \mathbf{h}_{i,\ell}$$

그런데 ELMo는,

- LM Pretrain 단계와 Downstream Task train 단계가 분리되어 있다.
- 그래서 LM이 task-agnostic하지만 이게 과연 장점일까?
- 이걸 하나로 합치는 건 안되나?
- 만약 합친다면 LM이 task를 더 정확하게 해결할 수 있는 보다 정교한 representation 역할을 할 수 있지 않을까?

CVT : Cross-View Training

- ELMo처럼 LM에만 의존하지 말고 최적화해야 할 task에 대해 지도학습을 하되, 부가적인 비지도학습 task를 통해, 이 비지도학습 task가 다양한 task 해결을 잘 따라 배우도록 유도하면 **Multi-task Learning**에 근접하게 되지 않을까?

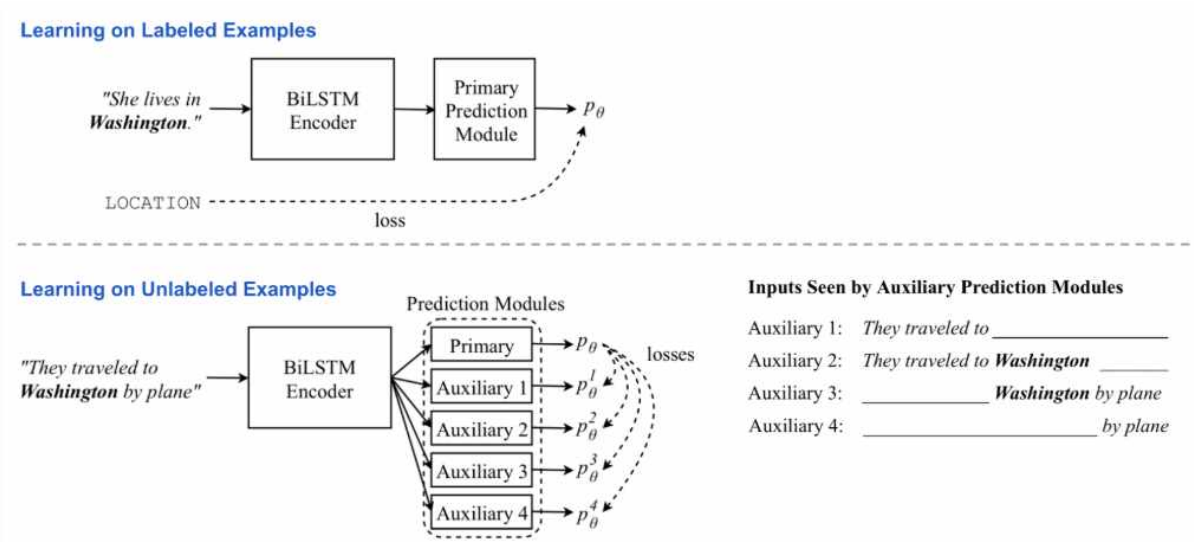
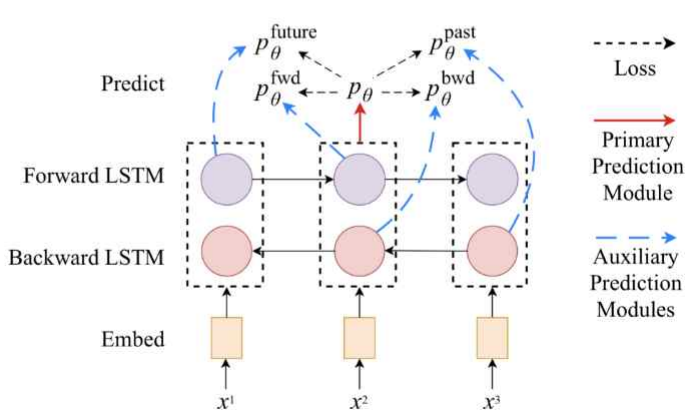


Fig. 4. The overview of semi-supervised language model cross-view training. (Image source: original paper)

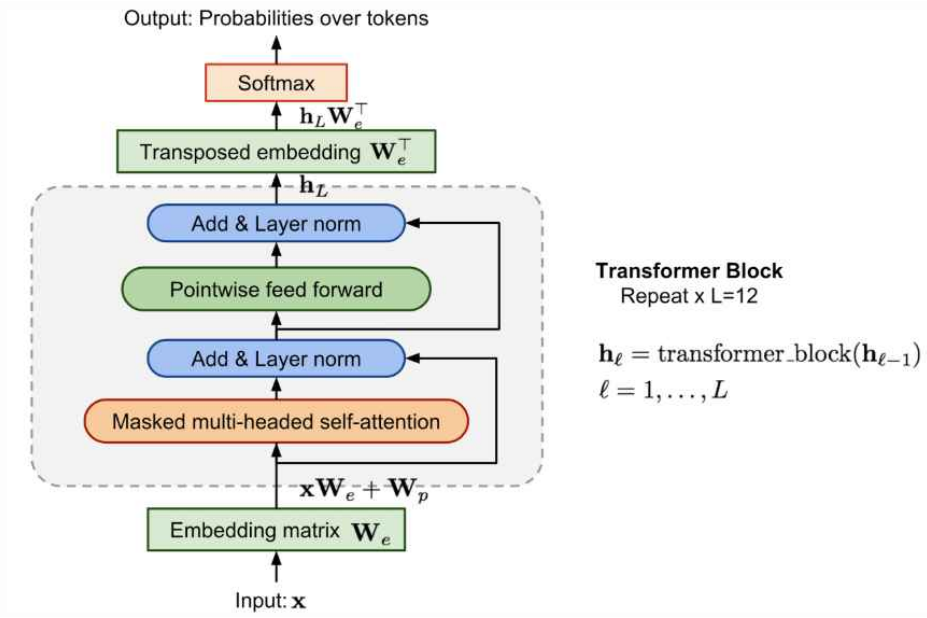
$$p_\theta^{\text{fwd}}(y_i \mid \mathbf{x}_i) = \text{NN}^{\text{fwd}}(\overrightarrow{\mathbf{h}}^{(i)})$$
$$p_\theta^{\text{bwd}}(y_i \mid \mathbf{x}_i) = \text{NN}^{\text{bwd}}(\overleftarrow{\mathbf{h}}^{(i)})$$
$$p_\theta^{\text{future}}(y_i \mid \mathbf{x}_i) = \text{NN}^{\text{future}}(\overrightarrow{\mathbf{h}}^{(i-1)})$$
$$p_\theta^{\text{past}}(y_i \mid \mathbf{x}_i) = \text{NN}^{\text{past}}(\overleftarrow{\mathbf{h}}^{(i+1)})$$



Task-agnostic한 LM을 만들려면 어떻게?

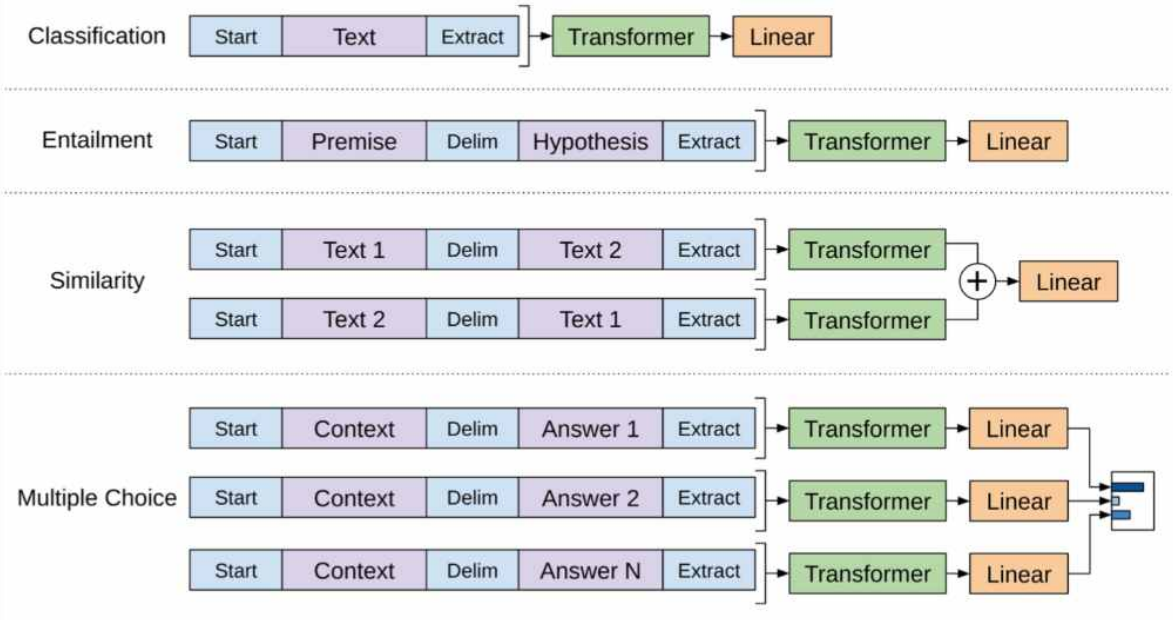
- Generative Pretrained LM이 필요하다.
- 즉, Decoder 같은 형태의 LM을 만들어야겠다.
- 그게 아니면 (지금까지 늘 그래왔듯이) task specific classifier를 optimize하는 식으로 트레이닝을 할 수밖에 없다.
- Transformer Decoder를 사용하면 어떨까?
  - 다음 토큰이 뭐가 올지 확률 모델을 잘 구현함

$$\mathcal{L}_{\text{LM}} = -\sum_i \log p(x_i \mid x_{i-k}, \ldots, x_{i-1})$$



GPT : Generative Pre-training Transformer

- 어떤 유형의 task도 Transformer에 입력되는 sentence로 치환하고, general한 Transformer Decoder 및 task별 classifier를 결합하여 한꺼번에 train한다. (ELMo보다 CVT 개념에 더 가깝다)
- 이후 pretrain된 Transformer 를 재활용하여 다른 task에 fine-tuning해도 잘 된다.



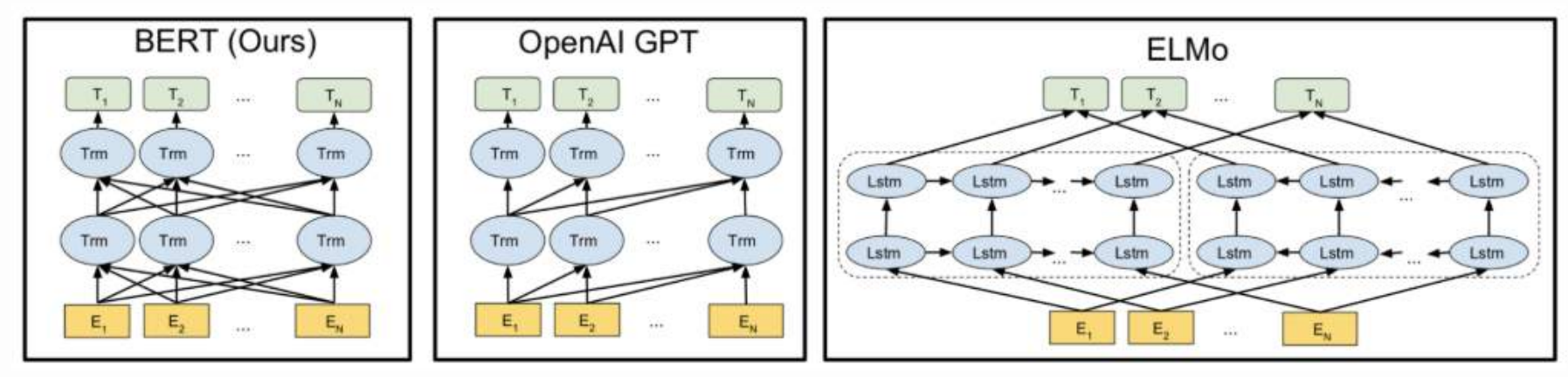
$$\mathcal{L}_{\text{cls}} = \sum_{(\mathbf{x},y)\in\mathcal{D}} \log P(y \mid x_1,\ldots,x_n) = \sum_{(\mathbf{x},y)\in\mathcal{D}} \log \text{softmax}(\mathbf{h}_L^{(n)}(\mathbf{x})\mathbf{W}_y)$$

$$\mathcal{L}_{\text{LM}} = -\sum_i \log p(x_i \mid x_{i-k},\ldots,x_{i-1})$$

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{LM}}$$

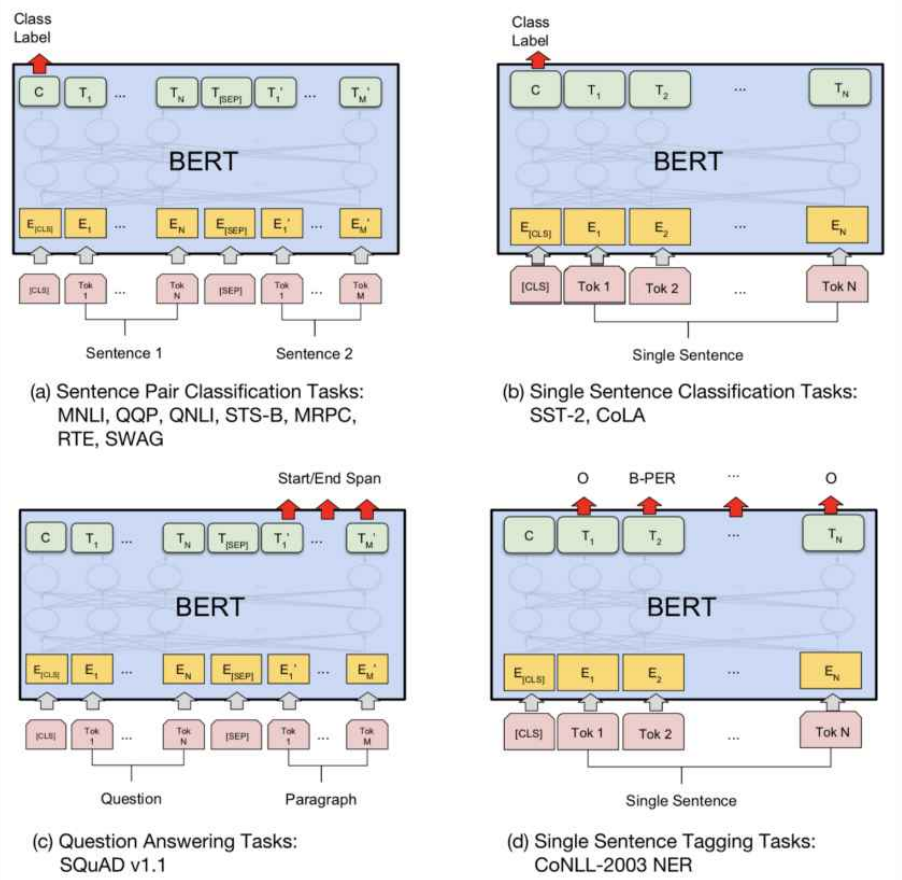
BERT : Bidirectional Encoder Representations from Transformers

- GPT : Transformer Decoder
- BERT : Transformer Encoder
- 여기에는 Language Model에 대한 관점의 차이가 숨어있다.
- BERT는 Attention 개념에 충실하게 LM을 구현하고 싶어한다.



BERT의 트레이닝

- Decoder가 아니라 Encoder이기 때문에 , LM의 구현과 최적화도 좀 다른 아이디어로 진행된다.
  - Task 1: Mask language model (MLM)
    - 일종의 Cloze Test 형식으로 구현
  - Task 2 : Next sentence prediction
    - Auxiliary task를 추가하여, 다음 문장이 무엇이 올까를 generation하기보다는 classfication하는 모델을 구현 (GPT라면 이런 게 필요없다)
- Task에 맞는 Fine-Tuning 구현이 필요하다.



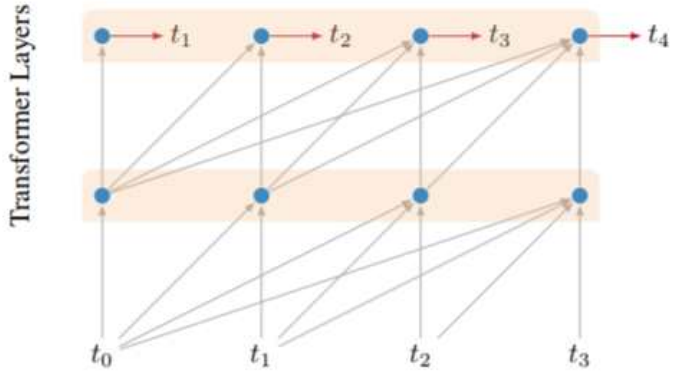
GPT vs. BERT 제 2라운드

- [Google] BERT -> Transformer-XL
- [OpenAI] GPT -> GPT-2

- 모델 구조의 큰 차이는 없고 모델 사이즈만 확 키웠다.
- 엄청 좋아졌다.

Character Level Modeling

- Al-Rfou et al.[2018] Character-Level Language Modeling with Deeper Self-Attention



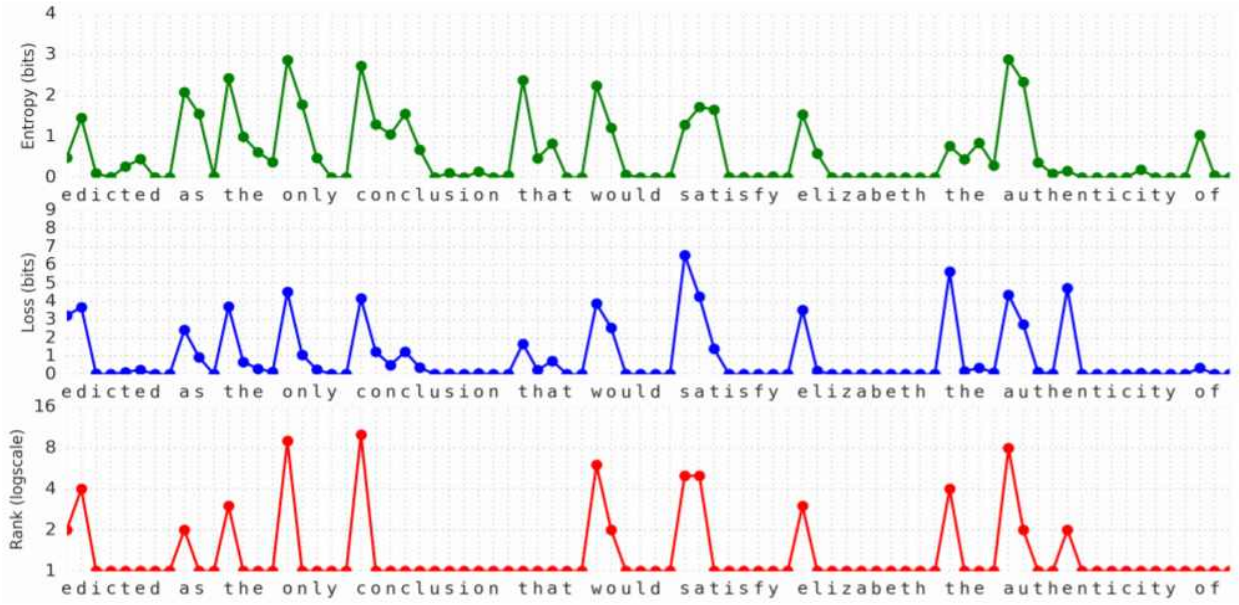


Figure 6: Per-character entropy, loss and rank assigned by T64 after seeding on the 512 character sequence from Figure 5.

Transformer-XL

- Segment-Level Recurrence with State Reuse
  - PPL 측면에서 BERT를 크게 개선하였다고 한다.

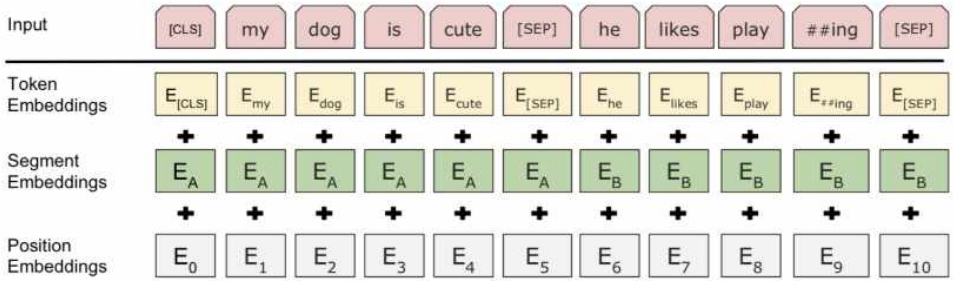


Fig. 11. BERT input representation. (Image source: original paper)

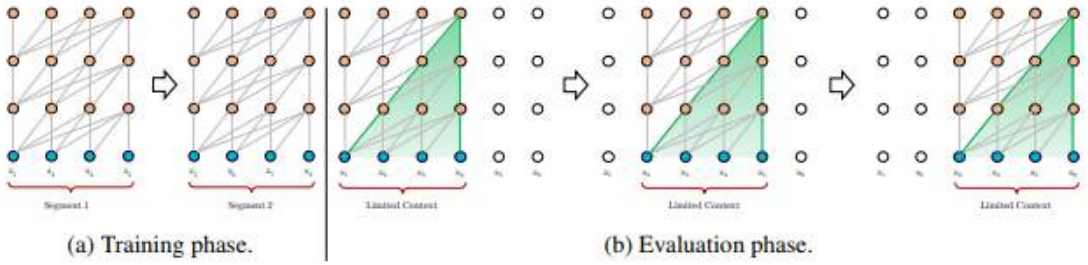


Figure 1: Illustration of the vanilla model with a segment length 4.

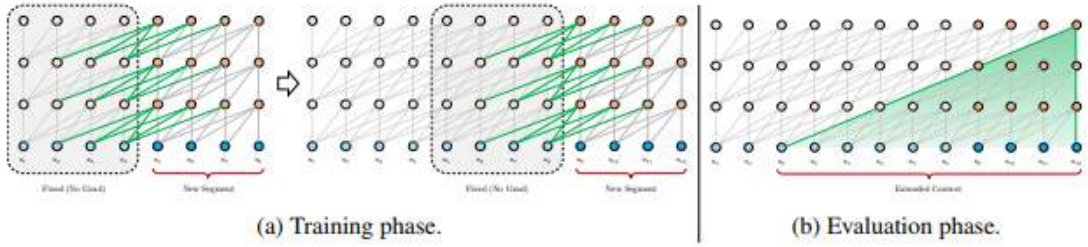


Figure 2: Illustration of the Transformer-XL model with a segment length 4.

GPT-2 : 덩치만 큰 GPT일까?

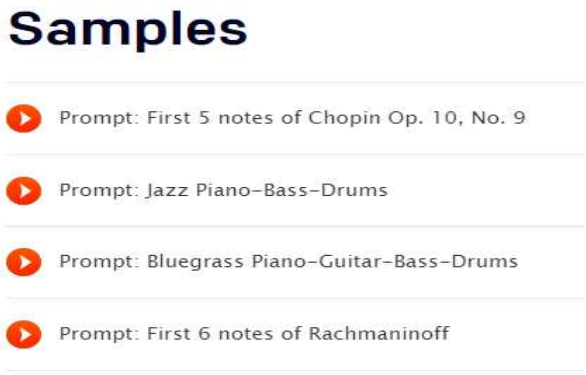
- 더욱 제너럴해진 GPT : no task-specific fine-tuning을 선언하다.
- Multitask Learning을 위한 문제 규정 관점의 전환

$$p(output|input) \longrightarrow p(output|input, task)$$

  - translation training
    (translate to french, english text, french text)
  - reading comprehension training
    (answer the question, document, question, answer)

MuseNet

- MuseNet uses the same general-purpose unsupervised technology as GPT-2, a large-scale transformer model trained to predict the next token in a sequence, whether audio or text.



https://openai.com/blog/musenet/