Data Mining
Quality Analytics

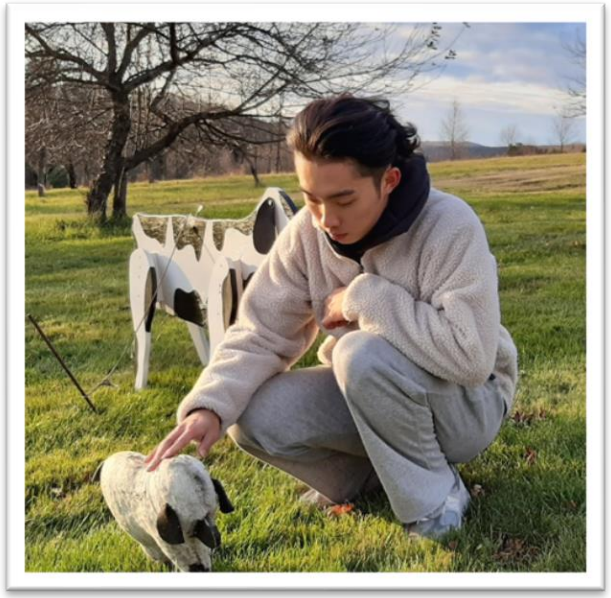
Open DMQA Seminar

# Beyond Bert

2020. 2. 5

Da Bin Min
Data Mining & Quality Analytics Lab.

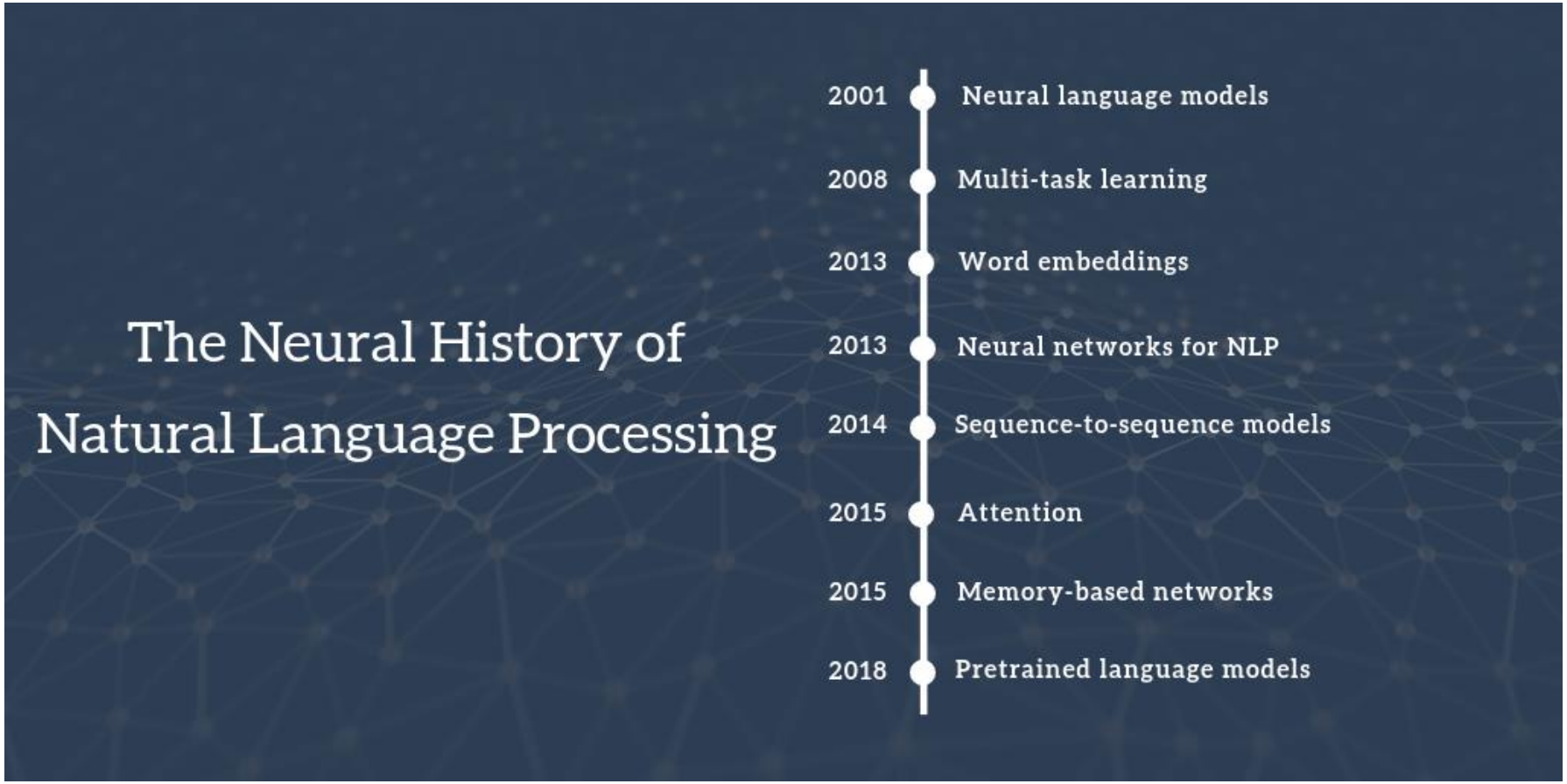
# Speaker



- 민다빈 (Dabin Min)
  - 고려대학교 산업경영공학부 재학 중
  - Data Mining & Quality Analytics Lab
  - 석사과정 (2019.03~)
- 관심 연구 분야
  - Machine Learning Algorithms
  - Knowledge Enhanced Language Model
- E-mail: reonaledo@korea.ac.kr

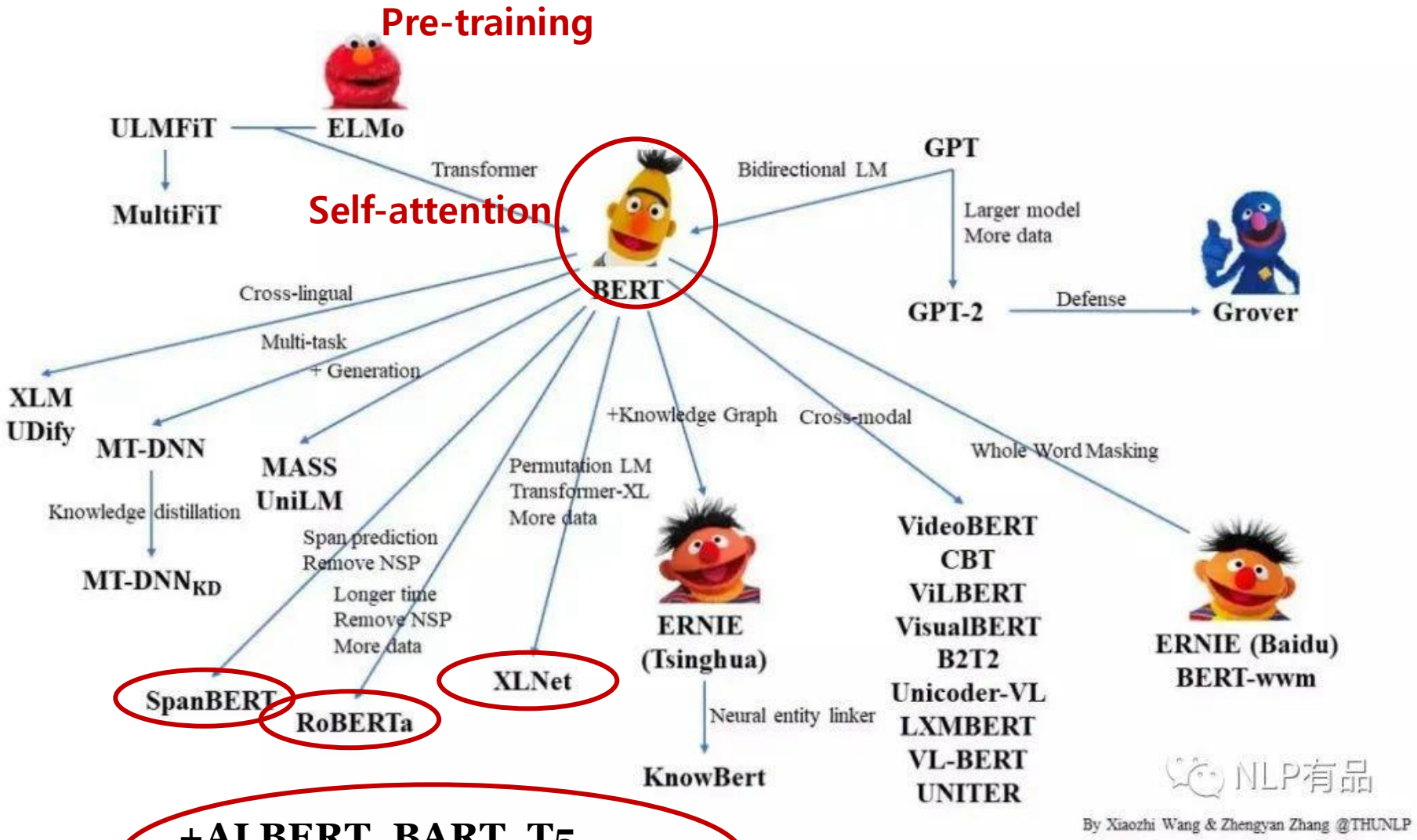
# Contents

# I. Introduction




https://ruder.io/a-review-of-the-recent-history-of-nlp/

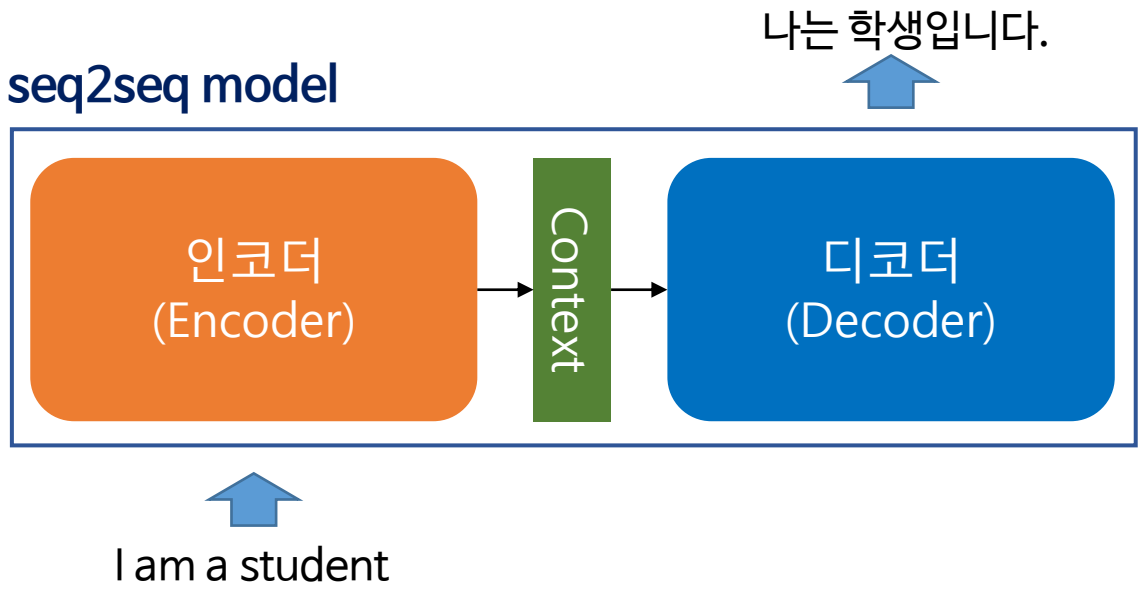
# I. Introduction

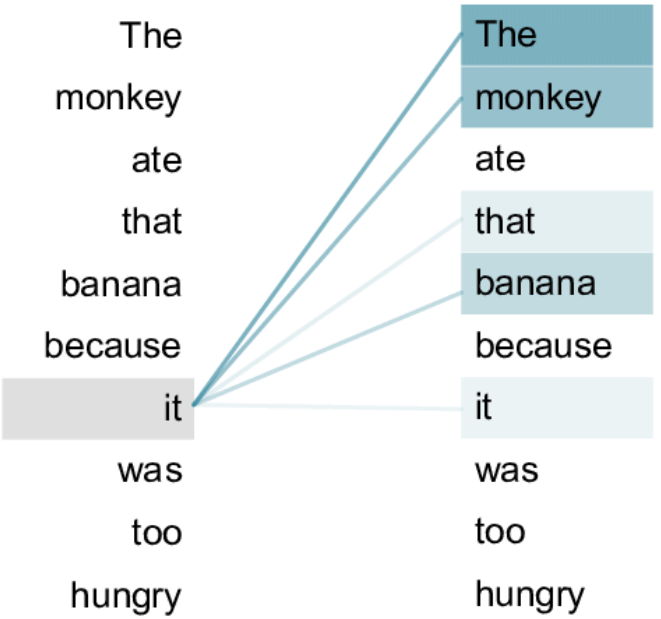



**+ALBERT, BART, T5, ELECTRA, GPT-3, DeBERTa**

# II. Remind: Transformer & BERT

- **Transformer**(2017. Google)
  - Sequence-to-sequence (e.g. Machine Translation)
  - 기존 seq2seq 모델의 한계
    - Long-term dependency problem
    - Parallelization
  - Self-attention
    - Bi-directional contextualized representation

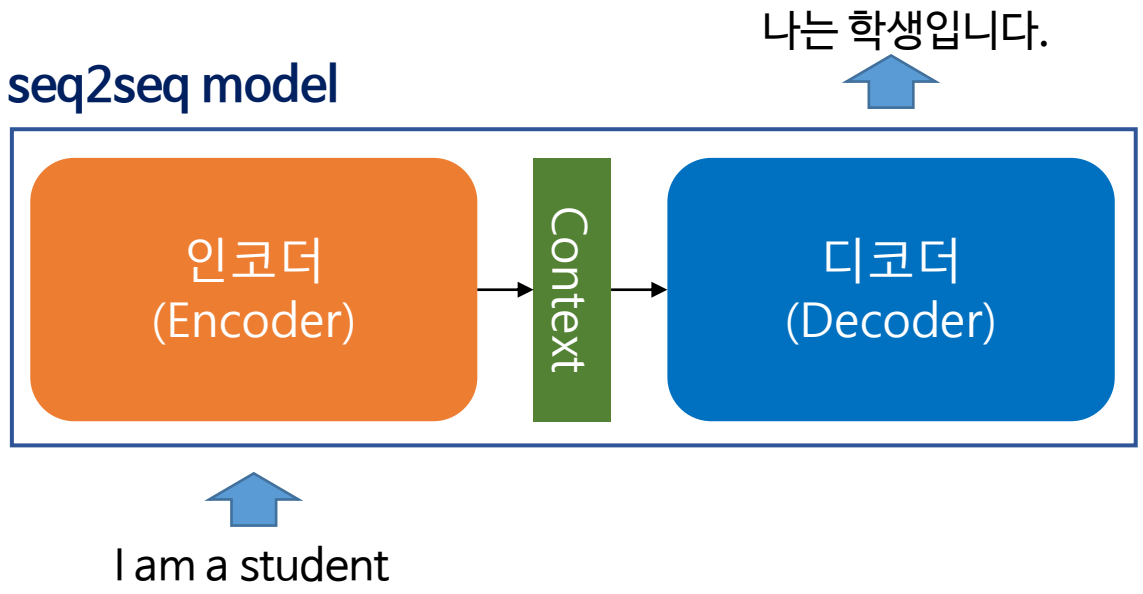


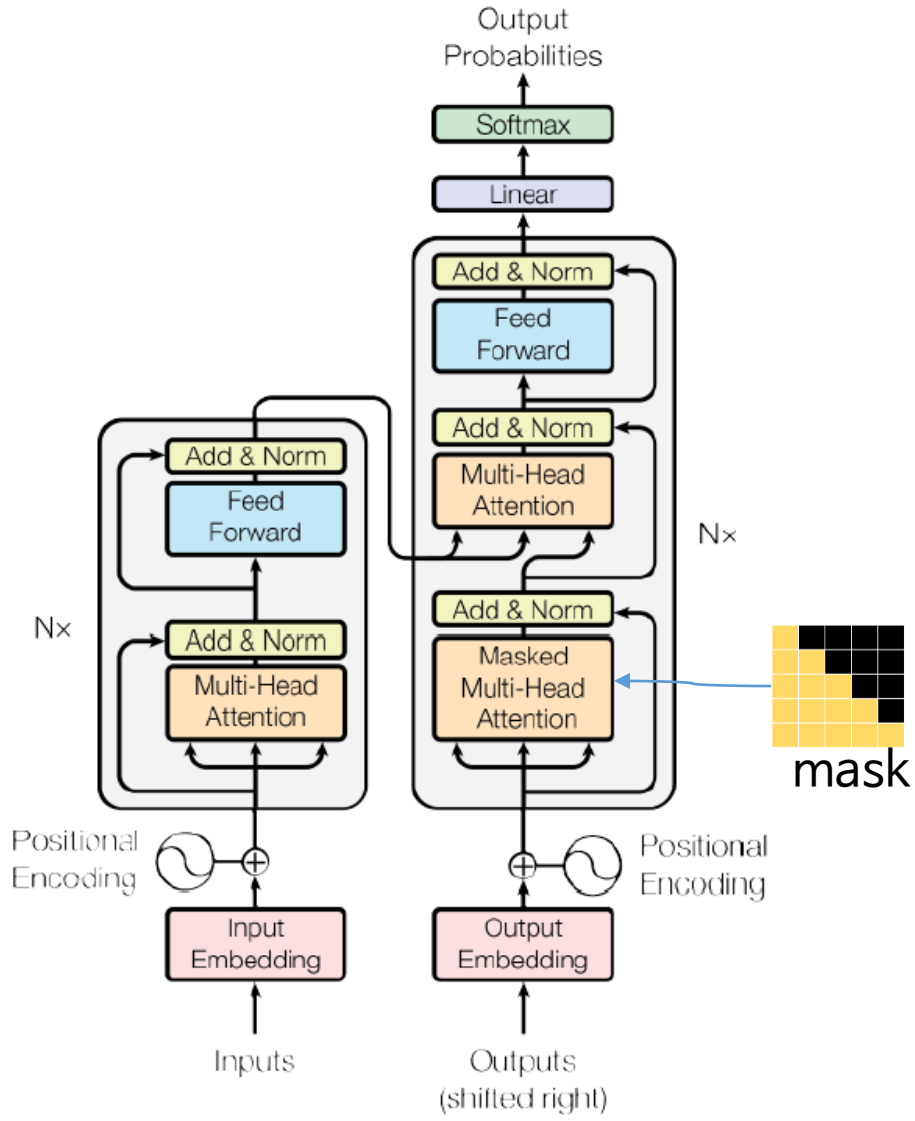




https://www.researchgate.net/figure/An-example-of-the-self-attention-mechanism-following-long-distance-dependency-in-the_fig1_342301870

# II. Remind: Transformer & BERT

- **Transformer**(2017 . Google)
  - Encoder-Decoder 구조
  - Positional encoding
  - Multi-head attention
  - Output masking



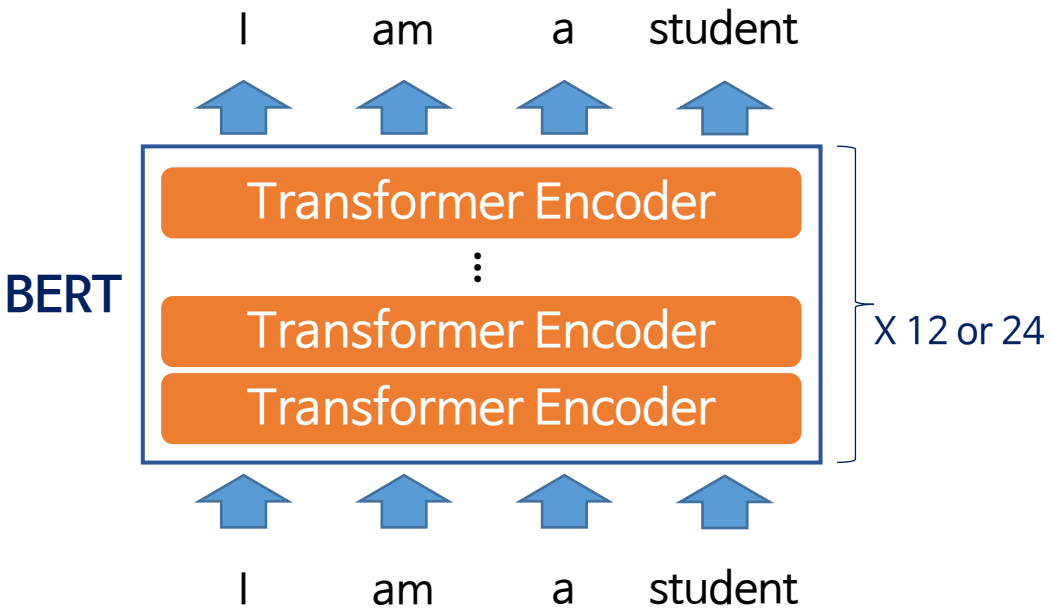



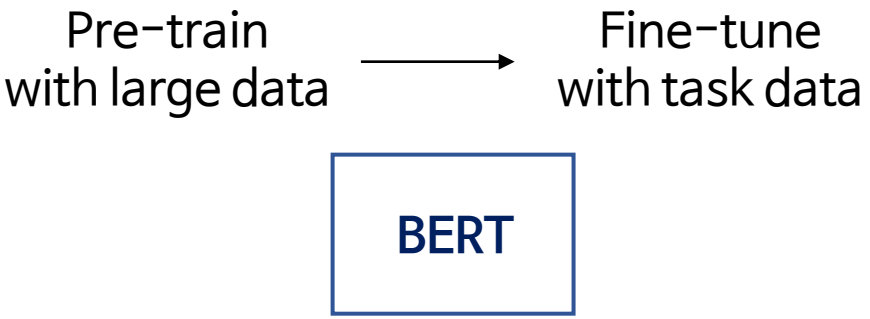



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In*Advances in Neural Information Processing Systems*(pp. 5998-6008).

# II. Remind: Transformer & BERT

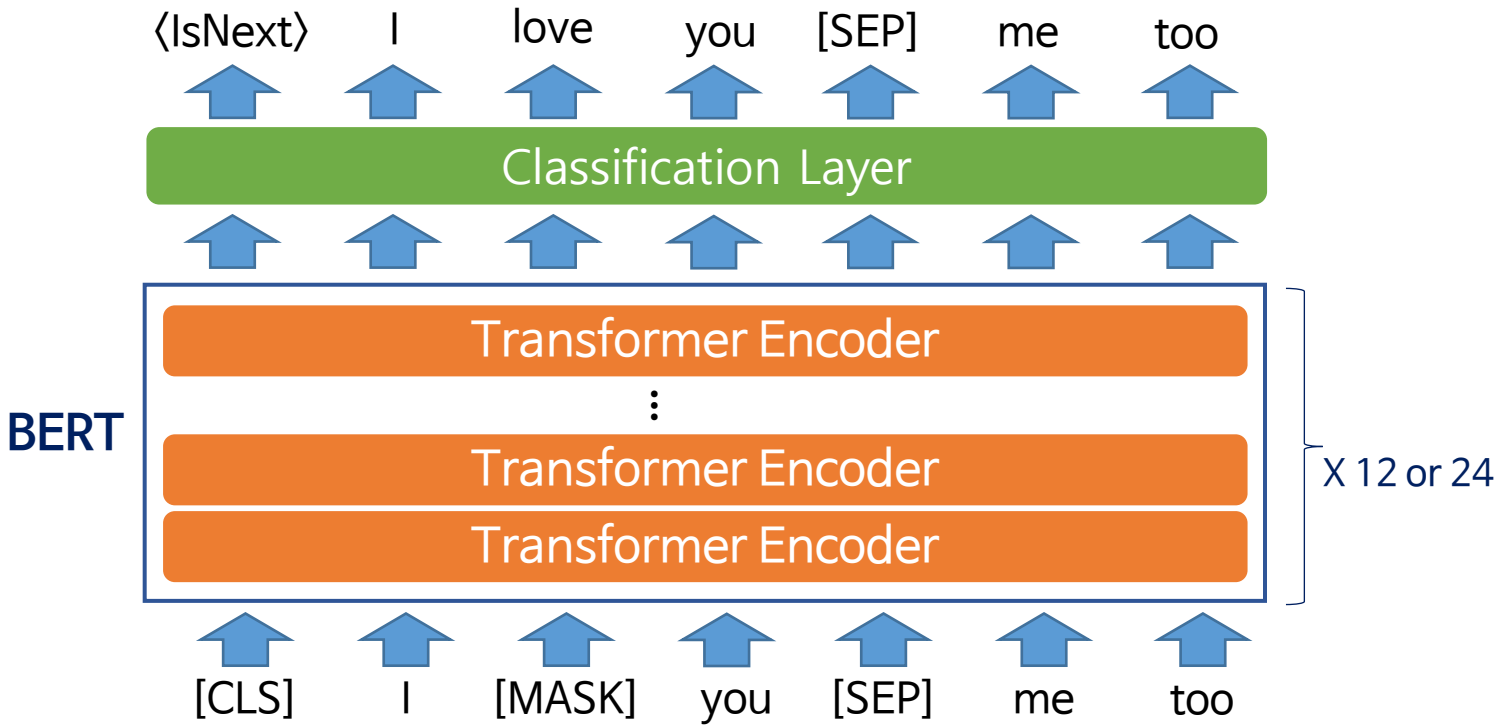
- **BERT** (Nov. 2018. Google)
  - Transformer encoder
  - Pre-training & fine-tuning



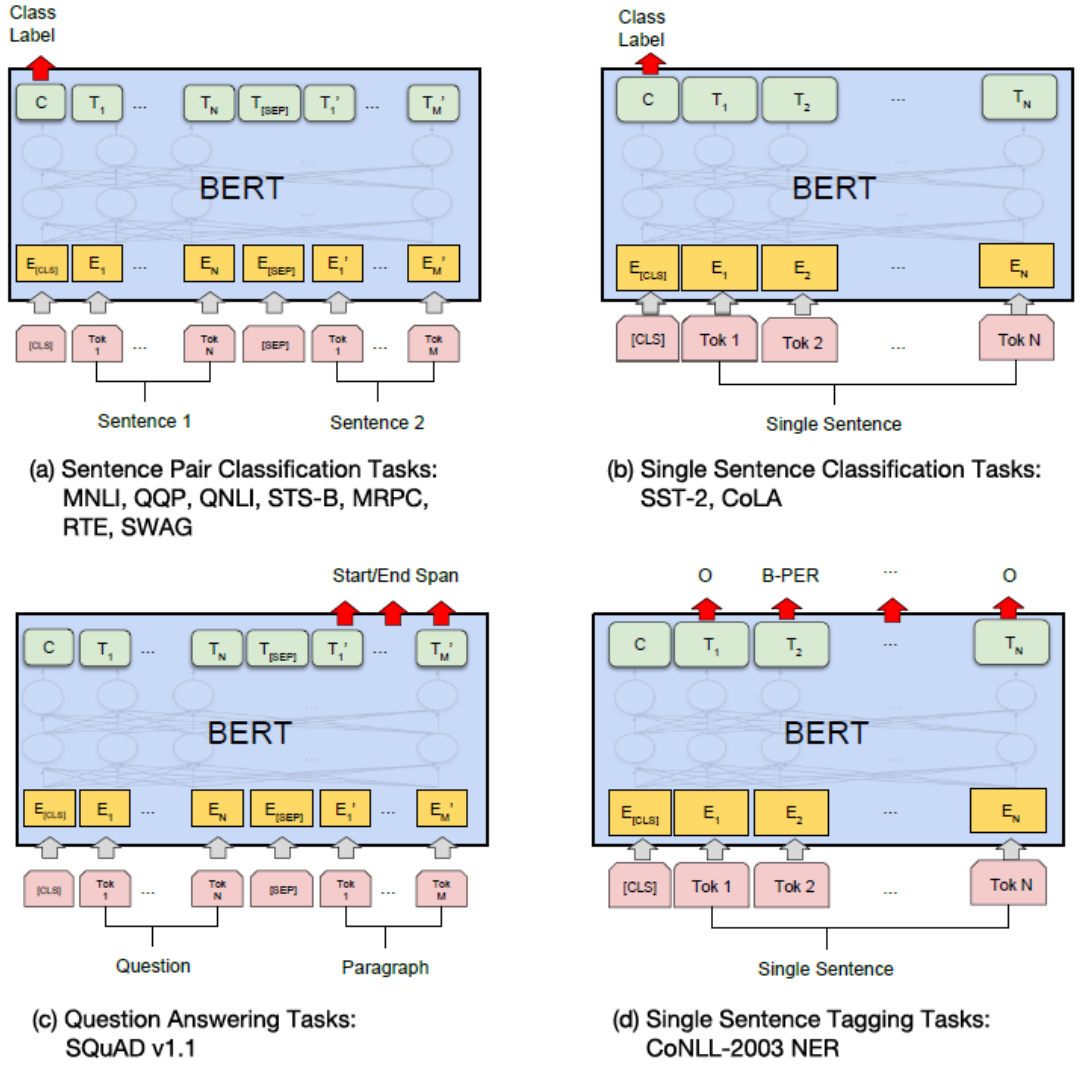
# II. Remind: Transformer & BERT

- **BERT** (Nov. 2018. Google)
  - Input 구성
    - 연속 또는 불연속한 2개의 segment (1개이상의 연속된 문장)
  - Pre-training task
    - Masked language modeling (MLM)
    - Next sentence prediction (NSP)



# II. Remind: Transformer & BERT

- **BERT** (Nov. 2018. Google)
  - Fine-tuning
    - Task 별로 입출력단 변경






J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. North Am. Association Computat. Linguistics (NAACL)-HLT, Minneapolis, MN, USA, June 2–7, 2019, pp. 4171–4186.

# III. 4 Ways to go beyond

1. **Pre-training method**
   - 사전훈련 방식 개선을 통한 성능 향상
   - SpanBERT, XLNet, RoBERTa, ALBERT, BART, ELECTRA, GPT-3, T5, DeBERTa
2. **Autoencoding (AE) + Autoregressive (AR)**
   - MLM을 통해 학습되는 BERT는 MASK 토큰 간 dependency를 학습할 수 없음
   - XLNet, BART, T5, DeBERTa-MT
3. **Model efficiency**
   - 더 적은 parameter, 더 적은 computation
   - ALBERT, ELECTRA
4. **Meta learning**
   - Generalized model, few-shot, zero-shot
   - GPT-3, T5

# III. 4 Ways to go beyond

1. **Pre-training method**
   - 사전훈련 방식 개선을 통한 성능 향상
   - SpanBERT, XLNet, RoBERTa, ALBERT, BART, ELECTRA, GPT-3, T5, DeBERTa
2. **Autoencoding (AE) + Autoregressive (AR)**
   - MLM을 통해 학습되는 BERT는 MASK 토큰 간 dependency를 학습할 수 없음
   - XLNet, BART, T5
3. **Model efficiency**
   - 더 적은 parameter, 단점은 computation
   - ALBERT, ELECTRA
4. **Meta learning**
   - Generalized model, few-shot, zero-shot
   - GPT-3, T5

**AutoEncoding(AE)**

전체 단어를 모두 보고 예측

$$X = [x_1, x_2, x_3, x_4, \dots, x_T]$$

$$\hat{X} = [x_1, [Mask], [Mask], x_4 \dots, x_T]$$

$$Likelihood: p(X|\hat{X}) \approx \prod_{t=1}^{T} p(x_t|\hat{X})$$

- Mask independence assumption
- Low text generation performance
- Pretrain-Finetune discrepancy

**AutoRegressive(AR)**

이전 단어들만 보고 예측

$$X = [x_1, x_2, x_3, x_4, \dots, x_T]$$

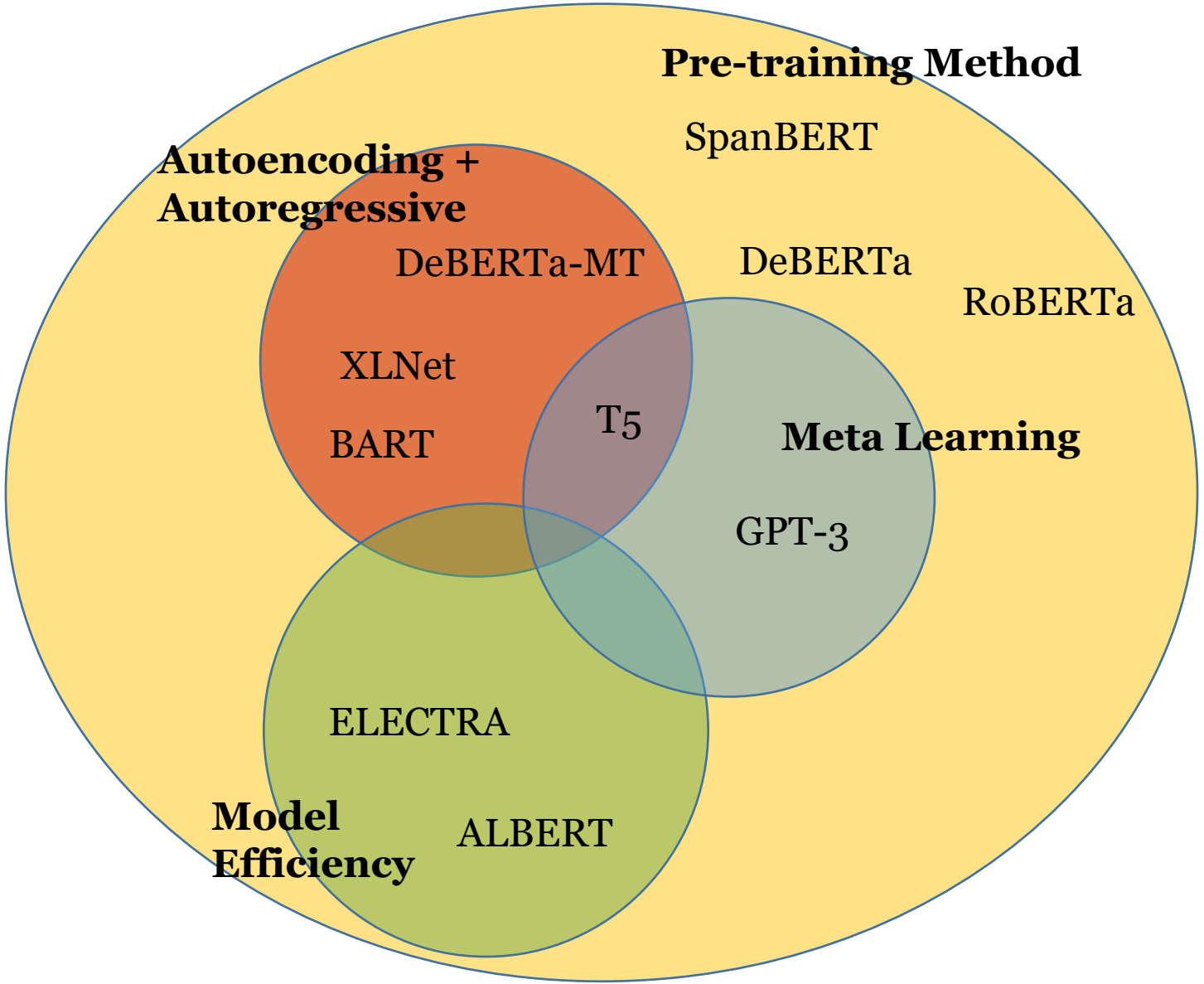
$$Likelihood: p(X) = \prod_{t=1}^{T} p(x_t|X_{<t})$$

- Uni-directional

# III. 4 Ways to go beyond

1. **Pre-training method**
   - 사전훈련 방식 개선을 통한 성능 향상
   - SpanBERT, XLNet, RoBERTa, ALBERT, BART, ELECTRA, GPT-3, T5, DeBERTa
2. **Autoencoding (AE) + Autoregressive (AR)**
   - MLM을 통해 학습되는 BERT는 MASK 토큰 간 dependency를 학습할 수 없음
   - XLNet, BART, T5, DeBERTa-MT
3. **Model efficiency**
   - 더 적은 parameter, 더 적은 computation cost
   - ALBERT, ELECTRA
4. **Meta learning**
   - Generalized model, few-shot, zero-shot
   - GPT-3, T5

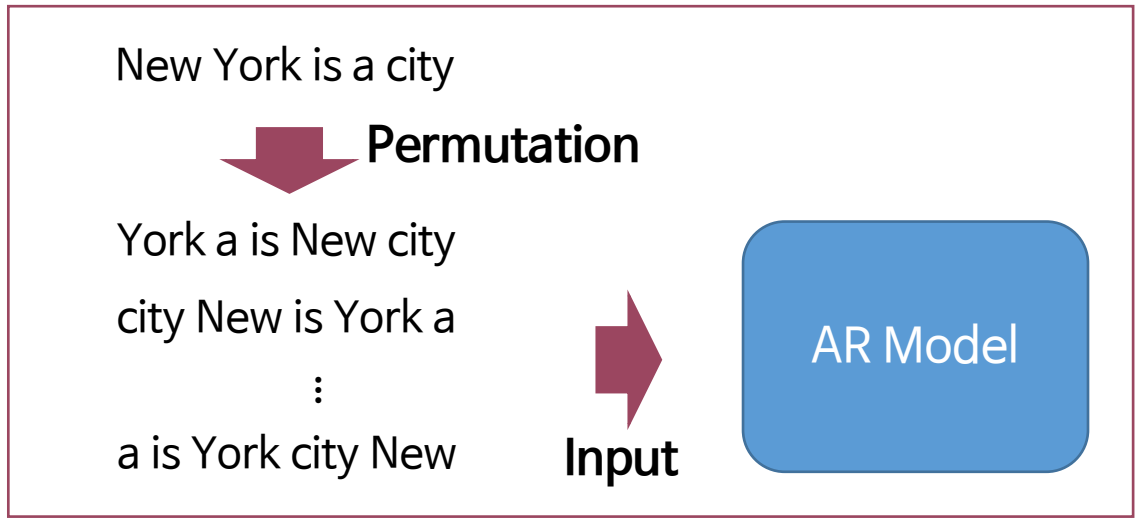
# III. 4 Ways to go beyond



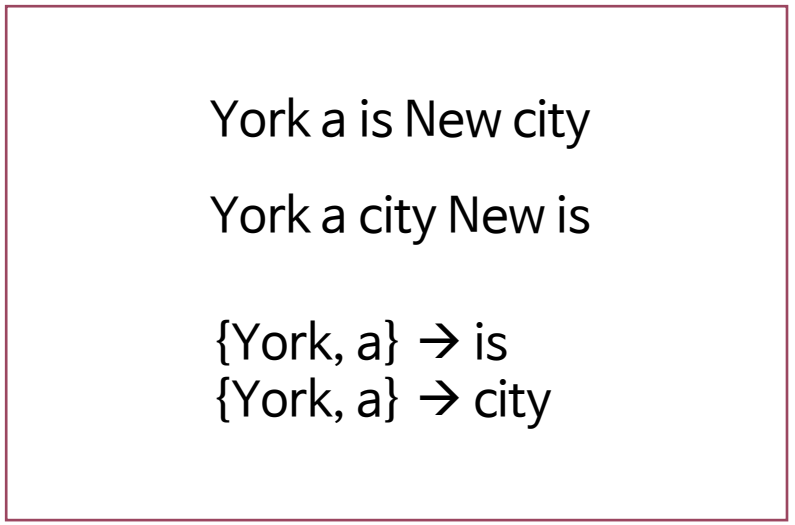
# IV. Models

1. **XLNet**(Jun. 2019. 카네기멜론대, Google)
   - AE 방식의 BERT의 한계점 지적 및 AE+AR 구조 제안
     - Permutation language modeling
     - Two-stream self-attention

**Permutation language modeling**




**Two-stream self-attention**

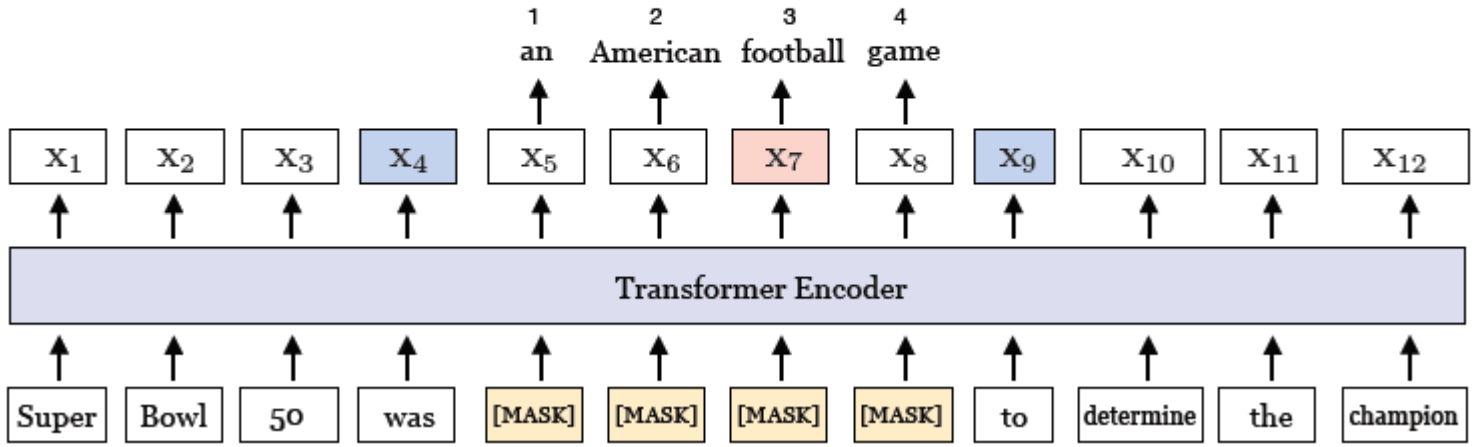


# IV. Models

## 2. SpanBERT (Jul. 2019, 워싱턴대, 프린스턴대, Allen AI Lab, Facebook )

- BERT의 Pre-training 방식 개선
  - Span masking
  - Span boundary objective
  - NSP 삭제, single sequence training

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$




Reprinted with from M. Joshi, "SpanBERT: Improving Pre-training by Representing and Predicting Spans," Accepted at TACL, 2020.

# IV. Models

3. **RoBERTa** (Jul. 2019. 워싱턴대, Facebook)
   - BERT는 underfitting 되었다 → BERT 최적화
     - Train longer, bigger batch, more data
     - NSP 삭제, single sequence training
     - Dynamic masking

| Model | data | bsz | steps | SQuAD (v1.1/2.0) | MNLI-m | SST-2 |
|---|---|---|---|---|---|---|
| RoBERTa | | | | | | |
| with BOOKS + WIKI | 16GB | 8K | 100K | 93.6/87.3 | 89.0 | 95.3 |
| + additional data (§3.2) | 160GB | 8K | 100K | 94.0/87.7 | 89.3 | 95.6 |
| + pretrain longer | 160GB | 8K | 300K | 94.4/88.7 | 90.0 | 96.1 |
| + pretrain even longer | 160GB | 8K | 500K | **94.6/89.4** | **90.2** | **96.4** |
| $\text{BERT}_{\text{LARGE}}$ | | | | | | |
| with BOOKS + WIKI | 13GB | 256 | 1M | 90.9/81.8 | 86.6 | 93.7 |
| $\text{XLNet}_{\text{LARGE}}$ | | | | | | |
| with BOOKS + WIKI | 13GB | 256 | 1M | 94.0/87.8 | 88.4 | 94.4 |
| + additional data | 126GB | 2K | 500K | 94.5/88.8 | 89.8 | 95.6 |

| | MNLI | QNLI | QQP | RTE | SST | MRPC | CoLA | STS | WNLI | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| *Single-task single models on dev* | | | | | | | | | | |
| $\text{BERT}_{\text{LARGE}}$ | 86.6/- | 92.3 | 91.3 | 70.4 | 93.2 | 88.0 | 60.6 | 90.0 | - | - |
| $\text{XLNet}_{\text{LARGE}}$ | 89.8/- | 93.9 | 91.8 | 83.8 | 95.6 | 89.2 | 63.6 | 91.8 | - | - |
| RoBERTa | **90.2/90.2** | **94.7** | **92.2** | **86.6** | **96.4** | **90.9** | **68.0** | **92.4** | **91.3** | - |
| *Ensembles on test (from leaderboard as of July 25, 2019)* | | | | | | | | | | |
| ALICE | 88.2/87.9 | 95.7 | **90.7** | 83.5 | 95.2 | 92.6 | **68.6** | 91.1 | 80.8 | 86.3 |
| MT-DNN | 87.9/87.4 | 96.0 | 89.9 | 86.3 | 96.5 | 92.7 | 68.4 | 91.1 | 89.0 | 87.6 |
| XLNet | 90.2/89.8 | 98.6 | 90.3 | 86.3 | **96.8** | **93.0** | 67.8 | 91.6 | **90.4** | 88.4 |
| RoBERTa | **90.8/90.2** | **98.9** | 90.2 | **88.2** | 96.7 | 92.3 | 67.8 | **92.2** | 89.0 | **88.5** |



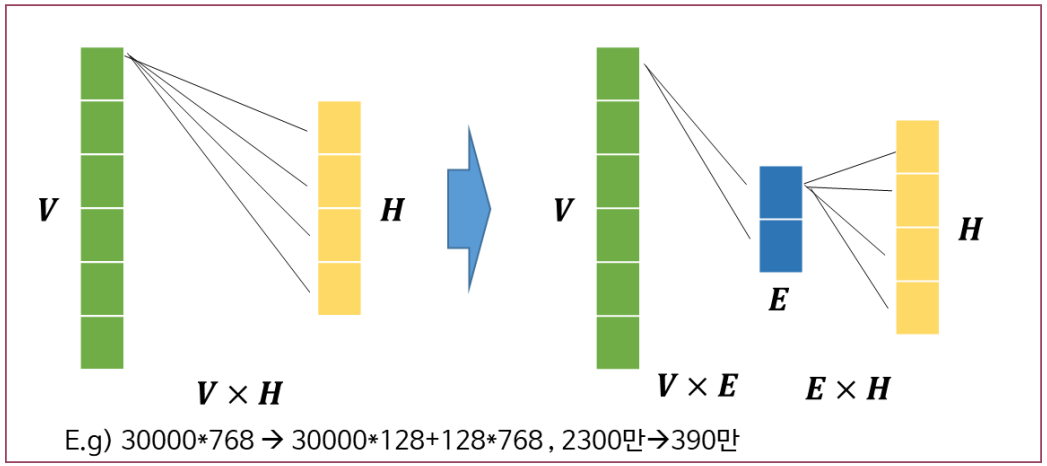
Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

# IV. Models

## 4. ALBERT (Sep. 2019. Google)

- BERT를 더 크게 만들기 위해 모델 효율화
  - Factorized embedding parameterization
  - Cross-layer parameter sharing
  - Sentence order prediction (SOP)

**Factorized Embedding Parameterization**




E.g) 30000*768 → 30000*128+128*768, 2300만→390만

| Model | | Parameters | SQuAD1.1 | SQuAD2.0 | MNLI | SST-2 | RACE | Avg | Speedup |
|---|---|---|---|---|---|---|---|---|---|
| BERT | base | 108M | 90.4/83.2 | 80.4/77.6 | 84.5 | 92.8 | 68.2 | 82.3 | 4.7x |
| | large | 334M | 92.2/85.5 | 85.0/82.2 | 86.6 | 93.0 | 73.9 | 85.2 | 1.0 |
| ALBERT | base | 12M | 89.3/82.3 | 80.0/77.1 | 81.6 | 90.3 | 64.0 | 80.1 | 5.6x |
| | large | 18M | 90.6/83.9 | 82.3/79.4 | 83.5 | 91.7 | 68.5 | 82.4 | 1.7x |
| | xlarge | 60M | 92.5/86.1 | 86.1/83.1 | 86.4 | 92.4 | 74.8 | 85.5 | 0.6x |
| | xxlarge | 235M | **94.1/88.3** | **88.1/85.1** | **88.0** | **95.2** | **82.3** | **88.7** | 0.3x |

| Models | MNLI | QNLI | QQP | RTE | SST | MRPC | CoLA | STS | WNLI | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| *Single-task single models on dev* | | | | | | | | | | |
| BERT-large | 86.6 | 92.3 | 91.3 | 70.4 | 93.2 | 88.0 | 60.6 | 90.0 | - | - |
| XLNet-large | 89.8 | 93.9 | 91.8 | 83.8 | 95.6 | 89.2 | 63.6 | 91.8 | - | - |
| RoBERTa-large | 90.2 | 94.7 | **92.2** | 86.6 | 96.4 | **90.9** | 68.0 | 92.4 | - | - |
| ALBERT (1M) | 90.4 | 95.2 | 92.0 | 88.1 | 96.8 | 90.2 | 68.7 | 92.7 | - | - |
| ALBERT (1.5M) | **90.8** | **95.3** | **92.2** | **89.2** | **96.9** | **90.9** | **71.4** | **93.0** | - | - |
| *Ensembles on test (from leaderboard as of Sept. 16, 2019)* | | | | | | | | | | |
| ALICE | 88.2 | 95.7 | **90.7** | 83.5 | 95.2 | 92.6 | **69.2** | 91.1 | 80.8 | 87.0 |
| MT-DNN | 87.9 | 96.0 | 89.9 | 86.3 | 96.5 | 92.7 | 68.4 | 91.1 | 89.0 | 87.6 |
| XLNet | 90.2 | 98.6 | 90.3 | 86.3 | 96.8 | 93.0 | 67.8 | 91.6 | 90.4 | 88.4 |
| RoBERTa | 90.8 | 98.9 | 90.2 | 88.2 | 96.7 | 92.3 | 67.8 | 92.2 | 89.0 | 88.5 |
| Adv-RoBERTa | 91.1 | 98.8 | 90.3 | 88.7 | 96.8 | 93.1 | 68.0 | 92.4 | 89.0 | 88.8 |
| ALBERT | **91.3** | **99.2** | 90.5 | **89.2** | **97.1** | **93.4** | 69.1 | **92.5** | **91.8** | **89.4** |

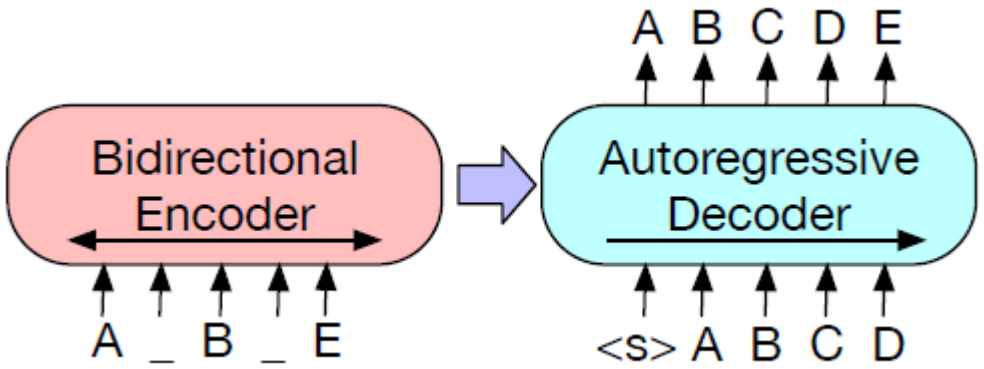


Z. Lan et al., "ALBERT: A Lite BERT for Selfsupervised Learning of Language Representations," in Int. Conf. Learning Representations, Addis Ababa, Ethiopia, May 2020.

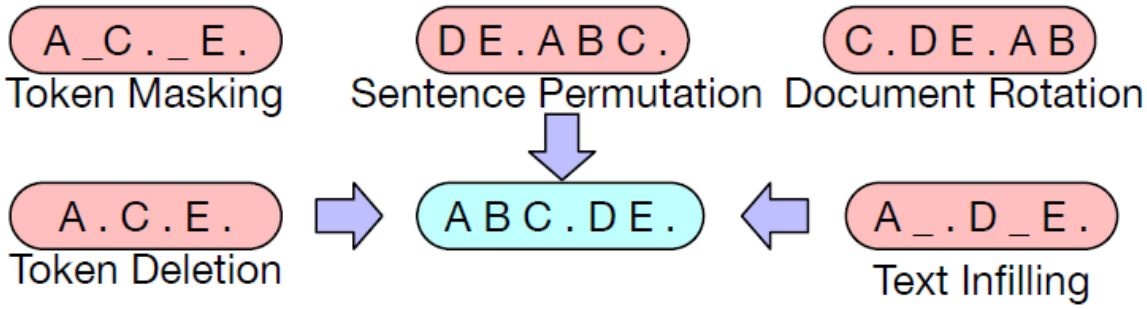
# IV. Models

5. **BART** (Oct. 2019. Facebook)
   - 새로운 AE+AR 구조 제안
     - Encoder와 decoder로 구성된 Transformer의 구조를 거의 그대로 이용(ReLU→GeLUs)
     - Noise flexibility

**BART Architecture**




**Noise flexibility**



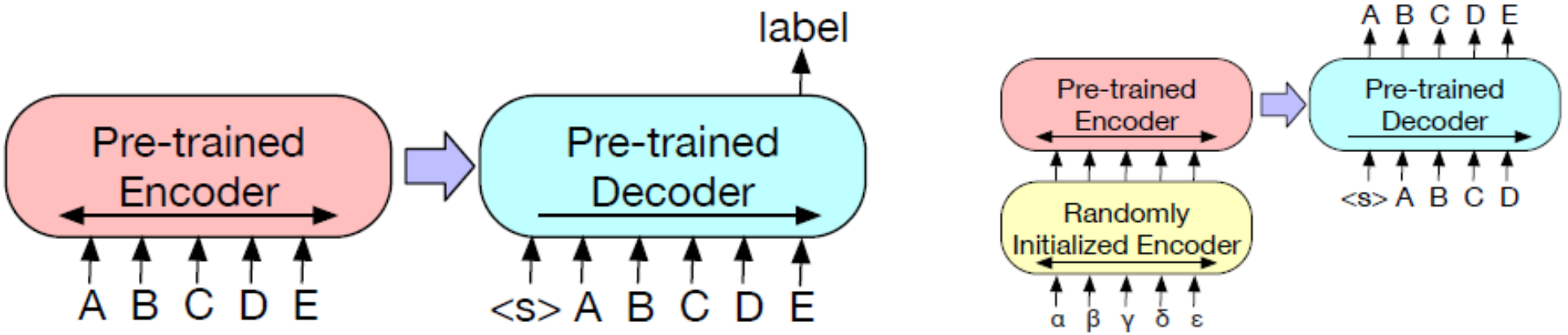



M. Lewis et al., “Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension.” arXiv preprint arXiv:1910.13461, 2019.

# IV. Models

5. **BART** (Oct. 2019. Facebook)
   - 새로운 AE+AR 구조 제안
     - Encoder와 decoder로 구성된 Transformer의 구조를 거의 그대로 이용(ReLU→GeLUs)
     - Noise flexibility

**Fine-tuning for classification & translation**






M. Lewis et al., "Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461, 2019.

# IV. Models

## 5. BART (Oct. 2019. Facebook)

- 새로운 AE+AR 구조 제안
  - Encoder와 decoder로 구성된 Transformer의 구조를 거의 그대로 이용(ReLU→GeLUs)
  - Noise flexibility

### 원문 뉴스

SK텔레콤(대표 박정호)이 자체 개발한 데이터센터용 AI 반도체를 25일 공개했다. 2024년 약 50조원 규모로 성장이 예상되는 AI 반도체 시장에 본격 진출한다는 계획이다. AI 반도체는 인공지능 서비스의 구현에 필요한 대규모 연산을 초고속·저전력으로 실행하는 비메모리 반도체다. 최근 데이터양이 기하급수적으로 늘면서 이를 처리하기 위한 AI 반도체의 필요성도 높아지고 있다. 현재 이 시장은 엔비디아·인텔·구글 등 글로벌 빅테크 기업 중심으로 경쟁이 치열해지고 있다. SK텔레콤은 이날 AI 반도체 'SAPEON X220'를 선보였다. 기존 GPU(그래픽처리장치) 대비 딥러닝 연산 속도가 1.5배 빠르고, 데이터센터에 적용 시 데이터 처리 용량이 1.5배 증가한다. 동시에 가격은 GPU의 절반 수준이고 전력 사용량도 80%에 불과하다. SK텔레콤은 맞춤형 설계를 통해 'SAPEON X220'의 경쟁력을 확보했다고 강조했다. 데이터 처리 역량 대부분을 동시다발적 데이터 처리에 활용하도록 설계해 효율성을 극대화했다는 설명이다. 이 제품은 다양한 분야의 데이터센터에 즉시 적용 가능하다. SK텔레콤은 국내외 다양한 사업자를 대상으로 AI 반도체 사업을 본격 추진할 계획이다.

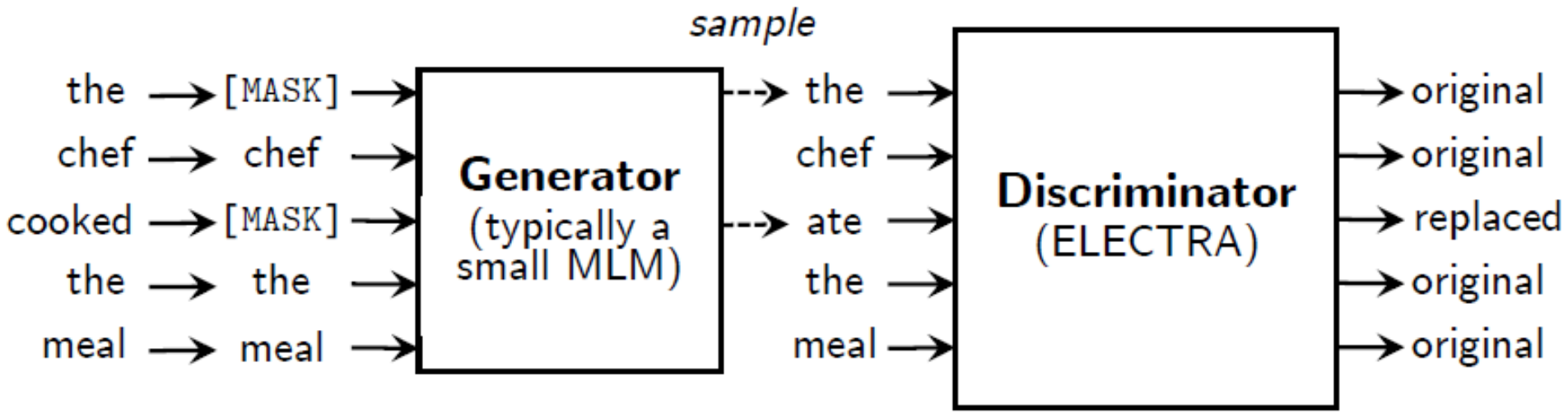
### 요약된 뉴스

25일 SK텔레콤(대표 박정호)이 자체 개발한 인공지능 서비스의 구현에 필요한 대규모 연산을 초고속·저전력으로 실행하는 비메모리 반도체인 AI 반도체를 선보이며, 글로벌 빅테크 기업 중심으로 경쟁이 치열해지고 있는 AI 반도체 시장에 본격 진출한다는 계획을 밝혔다.

https://github.com/SKT-AI/KoBART

# IV. Models

6. **ELECTRA** (Mar. 2020. 스탠포드, Google)
   - MLM 학습 시 계산 비효율성 지적
     - 오직 Masked token에 대해서만 학습이 이루어짐
   - GAN-like 구조 제안
     - Sample-efficient pre-training task




K. Clark et al., "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." in Int. Conf. Learning Representations, Addis Ababa, Ethiopia, May 2020.

# IV. Models

6. **ELECTRA** (Mar. 2020. 스탠포드, Google)
   - MLM 학습 시 계산 비효율성 지적
     - 오직 Masked token에 대해서만 학습이 이루어짐
   - GAN-like 구조 제안
     - Sample-efficient pre-training task

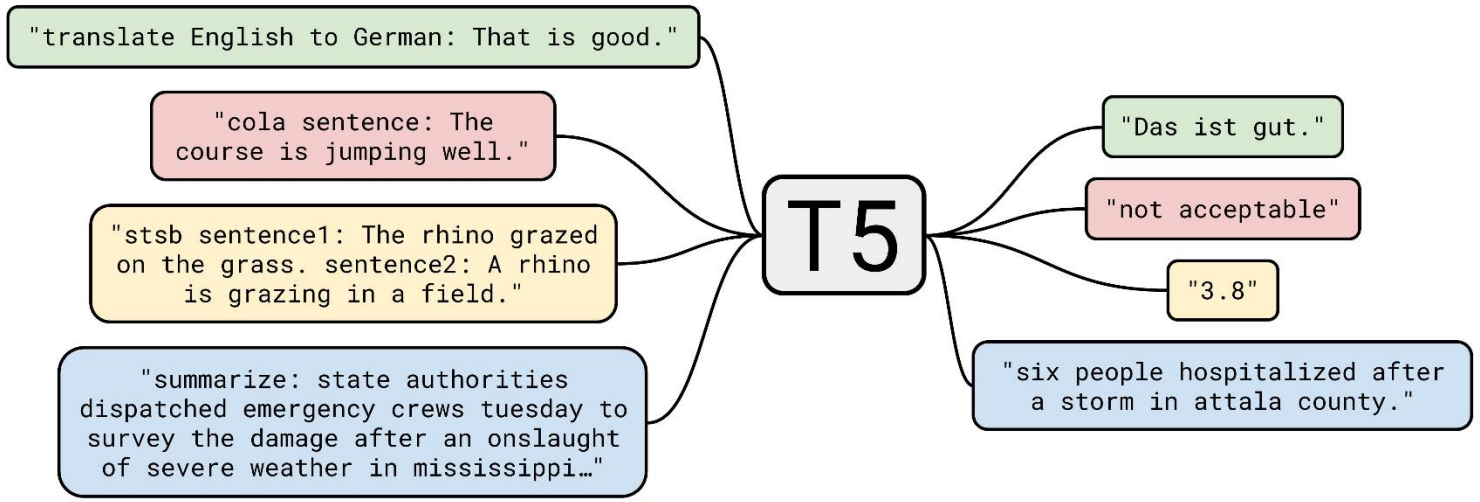
| Model | Train FLOPs | Params | CoLA | SST | MRPC | STS | QQP | MNLI | QNLI | RTE | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 1.9e20 (0.27x) | 335M | 60.6 | 93.2 | 88.0 | 90.0 | 91.3 | 86.6 | 92.3 | 70.4 | 84.0 |
| RoBERTa-100K | 6.4e20 (0.90x) | 356M | 66.1 | 95.6 | **91.4** | 92.2 | 92.0 | 89.3 | 94.0 | 82.7 | 87.9 |
| RoBERTa-500K | 3.2e21 (4.5x) | 356M | 68.0 | 96.4 | 90.9 | 92.1 | 92.2 | 90.2 | 94.7 | 86.6 | 88.9 |
| XLNet | 3.9e21 (5.4x) | 360M | 69.0 | **97.0** | 90.8 | 92.2 | 92.3 | 90.8 | 94.9 | 85.9 | 89.1 |
| BERT (ours) | 7.1e20 (1x) | 335M | 67.0 | 95.9 | 89.1 | 91.2 | 91.5 | 89.6 | 93.5 | 79.5 | 87.2 |
| ELECTRA-400K | 7.1e20 (1x) | 335M | **69.3** | 96.0 | 90.6 | 92.1 | **92.4** | 90.5 | 94.5 | 86.8 | 89.0 |
| ELECTRA-1.75M | 3.1e21 (4.4x) | 335M | 69.1 | 96.9 | 90.8 | **92.6** | **92.4** | **90.9** | **95.0** | **88.0** | **89.5** |

K. Clark et al., "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." in Int. Conf. Learning Representations, Addis Ababa, Ethiopia, May 2020.

# IV. Models

## 7. T5 & GPT-3 (Oct. 2019. Google / May. 2020. OpenAI)

- Fine-tuning 기반 모델의 한계 지적
  - Downstream task를 풀기 위해 많은 레이블 데이터 필요
  - Fine-tuning시 특정 task외의 문제에 대한 일반화 능력 상실
- Meta-learning 방식 채택
  - Task의 종류도 텍스트로 처리하여 함께 모델 인풋에 사용
  - T5: multitask learning, transformer encoder-decoder 구조, 110억개 파라메터
  - GPT-3: few-shot learning, transformer decoder x96,1750억개 파라메터



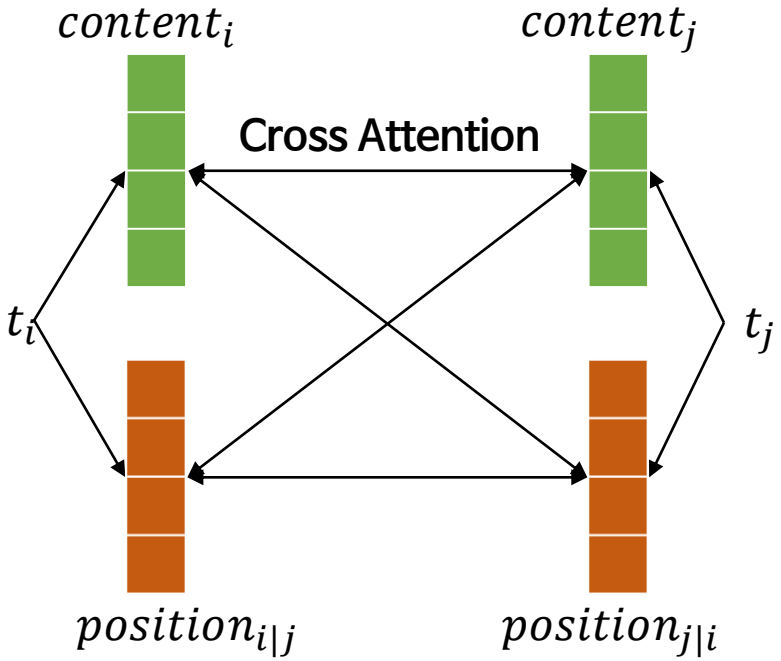



Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019

# IV. Models

8. **DeBERTa** (Jun. 2020. Microsoft)
   - BERT, RoBERTa 계열 발전 (21년 2월 기준)
     - Disentangled attention mechanism
     - Enhanced mask decoder
   - 기존 아이디어 채택
     - NSP 삭제 (RoBERTa, 2019)
     - AE + AR for generation task (UniLM, 2019)
     - Span masking (SpanBERT. 2019)

**Disentangled attention mechanism**




$$A_{i,j} = \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^T$$
$$= H_i H_j^T + H_i P_{j|i}^T + P_{i|j} H_j^T + P_{i|j} P_{j|i}^T$$

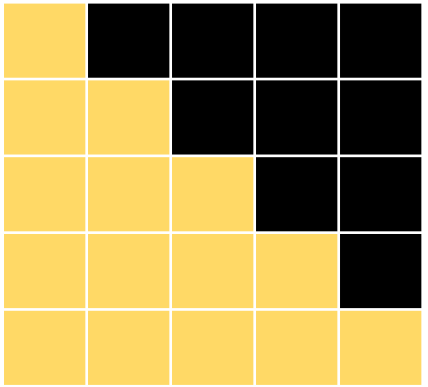
# IV. Models

8. **DeBERTa** (Jun. 2020. Microsoft)
   - BERT, RoBERTa 계열 발전 (21년 2월 기준)
     - Disentangled attention mechanism
     - Enhanced mask decoder
   - 기존 아이디어 채택
     - NSP 삭제 (RoBERTa, 2019)
     - AE + AR for generation task (UniLM, 2019)
     - Span masking (SpanBERT. 2019)

**AE+AR for Generation Task**



Attention mask for AE



Attention mask for AR

**DeBERTa-MT**

Pre-training: 한 batch에 일부는 AE, 일부는 AR 마스킹 적용

Fine-tuning: AR 마스킹만 적용

# V. Conclusion

## GLUE Benchmark

| | Rank | Name | Model | URL | Score | CoLA | SST-2 | MRPC | STS-B | QQP | MNL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | DeBERTa Team - Microsoft | DeBERTa / TuringNLRv4 | | 90.8 | 71.5 | 97.5 | 94.0/92.0 | 92.9/92.6 | 76.2/90.8 | |
| | 2 | HFL iFLYTEK | MacALBERT + DKM | | 90.7 | 74.8 | 97.0 | 94.5/92.6 | 92.8/92.6 | 74.7/90.6 | |
| + | 3 | Alibaba DAMO NLP | StructBERT + TAPT | | 90.6 | 75.3 | 97.3 | 93.9/91.9 | 93.2/92.7 | 74.8/91.0 | |
| + | 4 | PING-AN Omni-Sinitic | ALBERT + DAAF + NAS | | 90.6 | 73.5 | 97.2 | 94.0/92.0 | 93.0/92.4 | 76.1/91.0 | |
| | 5 | ERNIE Team - Baidu | ERNIE | | 90.4 | 74.4 | 97.5 | 93.5/91.4 | 93.0/92.6 | 75.2/90.9 | |
| | 6 | T5 Team - Google | T5 | | 90.3 | 71.6 | 97.5 | 92.8/90.4 | 93.1/92.8 | 75.1/90.6 | |
| | 7 | Microsoft D365 AI & MSR AI & GATECH | MT-DNN-SMART | | 89.9 | 69.5 | 97.5 | 93.7/91.6 | 92.9/92.5 | 73.9/90.2 | |
| + | 8 | Huawei Noah's Ark Lab | NEZHA-Large | | 89.8 | 71.7 | 97.3 | 93.3/91.0 | 92.4/91.9 | 75.2/90.7 | |
| + | 9 | Zihang Dai | Funnel-Transformer (Ensemble B10-10-10H1024) | | 89.7 | 70.5 | 97.5 | 93.4/91.2 | 92.6/92.3 | 75.4/90.7 | |
| + | 10 | ELECTRA Team | ELECTRA-Large + Standard Tricks | | 89.4 | 71.7 | 97.1 | 93.1/90.7 | 92.9/92.5 | 75.6/90.8 | |

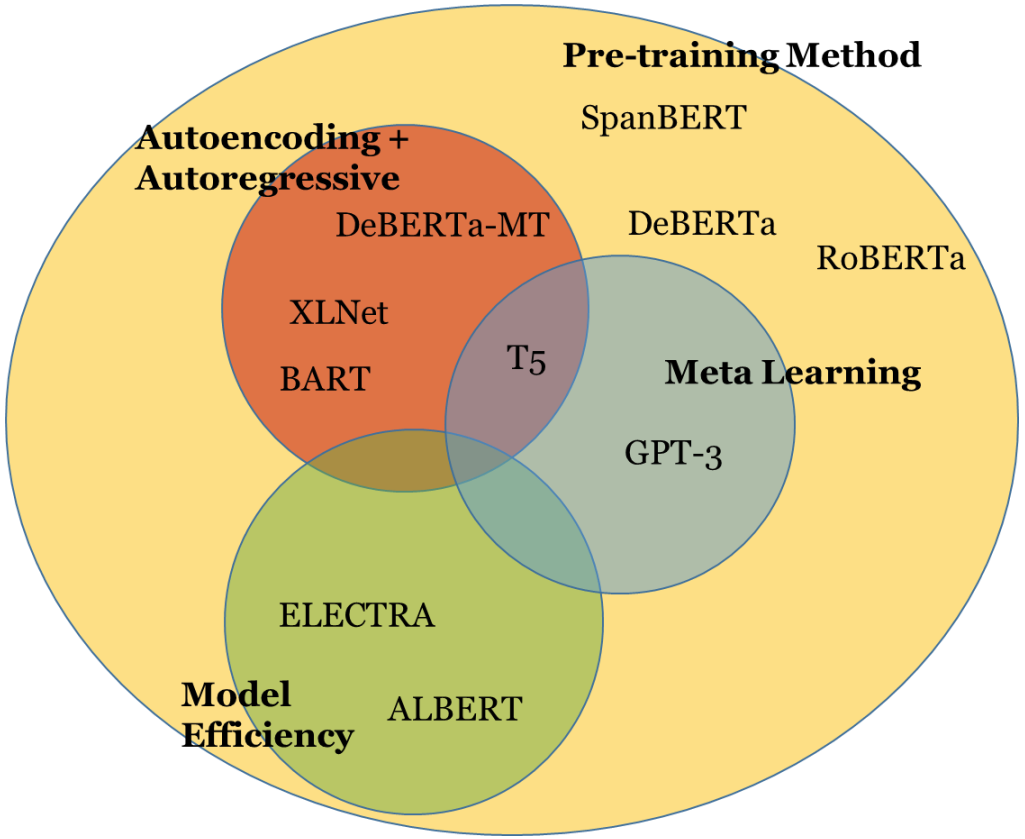
# V. Conclusion

## SuperGLUE Benchmark

| Rank | Name | Model | URL | Score | BoolQ | CB | COPA | MultiRC | ReCoRD | RTE | WiC | WSC | AX-b | AX-g |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | DeBERTa Team - Microsoft | DeBERTa / TuringNLRv4 | | 90.3 | 90.4 | 95.7/97.6 | 98.4 | 88.2/63.7 | 94.5/94.1 | 93.2 | 77.5 | 95.9 | 66.7 | 93.3/93.8 |
| 2 | Zirui Wang | T5 + Meena, Single Model (Meena Team - Google Brain) | | 90.2 | 91.3 | 95.8/97.6 | 97.4 | 88.3/63.0 | 94.2/93.5 | 92.7 | 77.9 | 95.9 | 66.5 | 88.8/89.9 |
| 3 | SuperGLUE Human Baselines | SuperGLUE Human Baselines | | 89.8 | 89.0 | 95.8/98.9 | 100.0 | 81.8/51.9 | 91.7/91.3 | 93.6 | 80.0 | 100.0 | 76.6 | 99.3/99.7 |
| 4 | T5 Team - Google | T5 | | 89.3 | 91.2 | 93.9/96.8 | 94.8 | 88.1/63.3 | 94.1/93.4 | 92.5 | 76.9 | 93.8 | 65.6 | 92.7/91.9 |
| 5 | Huawei Noah's Ark Lab | NEZHA-Plus | | 86.7 | 87.8 | 94.4/96.0 | 93.6 | 84.6/55.1 | 90.1/89.6 | 89.1 | 74.6 | 93.2 | 58.0 | 87.1/74.4 |
| 6 | Alibaba PAI&ICBU | PAI Albert | | 86.1 | 88.1 | 92.4/96.4 | 91.8 | 84.6/54.7 | 89.0/88.3 | 88.8 | 74.1 | 93.2 | 75.6 | 98.3/99.2 |
| 7 | Tencent Jarvis Lab | RoBERTa (ensemble) | | 85.9 | 88.2 | 92.5/95.6 | 90.8 | 84.4/53.4 | 91.5/91.0 | 87.9 | 74.1 | 91.8 | 57.6 | 89.3/75.6 |
| 8 | Zhuiyi Technology | RoBERTa-mtl-adv | | 85.7 | 87.1 | 92.4/95.6 | 91.2 | 85.1/54.3 | 91.7/91.3 | 88.1 | 72.1 | 91.8 | 58.5 | 91.0/78.1 |
| 9 | Infosys : DAWN : AI Research | RoBERTa-iCETS | | 85.0 | 86.2 | 93.2/95.2 | 91.2 | 84.6/53.4 | 89.9/89.3 | 88.5 | 72.1 | 89.0 | 35.2 | 93.8/68.8 |
| 10 | Facebook AI | RoBERTa | | 84.6 | 87.1 | 90.5/95.2 | 90.6 | 84.4/52.5 | 90.6/90.0 | 88.2 | 69.9 | 89.0 | 57.9 | 91.0/78.1 |

# V. Conclusion

- BERT 이후 등장한 연구들의 4가지 개선 방향
  - Pre-training method
  - Autoencoding + Autoregressive
  - Model efficiency
  - Meta learning



# 감사합니다

# 참고문헌

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.
[2] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. North Am. Association Computat. Linguistics (NAACL)-HLT, Minneapolis, MN, USA, June 2–7, 2019, pp. 4171–4186.
[3] M. Joshi et al., “SpanBERT: Improving pre-training by representing and predicting spans,” arXiv preprint 1907.10529, 2019.
[4] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pp. 5754–5764, 2019.
[5] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.
[6] Z. Lan et al., “ALBERT: A Lite BERT for Selfsupervised Learning of Language Representations,” in Int. Conf. Learning Representations, Addis Ababa, Ethiopia, May 2020.
[7] M. Lewis et al., “Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension.” arXiv preprint arXiv:1910.13461, 2019.
[8] K. Clark et al., “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.” in Int. Conf. Learning Representations, Addis Ababa, Ethiopia, May 2020.
[9] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint 2006.03654, 2020.
[10] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In Advances in Neural Information Processing Systems, pp. 13042–13054, 2019.