T5 and Large Language Model

timeline

Seq2Seq ⇒ Attention+Seq2Seq ⇒ Transformer

⇓

GPT-1 (Dec) ⇒ BERT (Enc)

⇓

GPT-2 ⇒ XLNET ⇒ RoBERTa ⇒ MASS

⇓

MASS ⇒ BART ⇒ MT-DNN ⇒ T5 (Enc-Dec)

모든 NLP task를 통합할 수 있도록 Text-to-text framework 사용

Pre-training (사전학습)의 대표적인 Objective (XLNET)

- Auto Encoding
- Auto Regressive

Auto Encoding (AE)

- BERT는 Denoising AE로 볼 수 있음

- Word sequence $\bar{x} = [x_1, x_2, \cdots, x_T]$
- Corrupted sequence $\hat{x} = [x_1, [MASK], \cdots, x_T]$ ← 복원할 자리
- likelihood $p(\bar{x}|\hat{x}) \approx \prod_{t=1}^{T} p(x_t|\hat{x})$
- Objective f. $\max_\theta \log p_\theta(\bar{x}|\hat{x})$

$$\approx \sum_{t=1}^{T} m_t \log p_\theta(x_t|\hat{x})$$

$$= \sum_{t=1}^{T} m_t \log \frac{\exp(H_\theta(\hat{x})_t^T e(x_t))}{\exp(H_\theta(\sum_{x'} \exp(H_\theta(\hat{x})_t^T e(x'))}$$

Auto Regressive (AR)

- 주로 decoder에서 사용, 특정 seq 주어지면 다음 seq를 한방향으로 예측

T5 Model

Pretrain → Finetune → Evaluation

1. text to text

input과 output이 text로 구성

text 형태로 주어진 문제에서 text 정답찾기

2. Transfer learning

BERT Style (encoder-only)은 Classification, Span prediction 특화

T5(encoder-decoder 구조)은 모든 NLP task에서 동일 모델, loss, hyper param 사용해도 잘 작동

- Enc-only, Dec-only 보다 basic transformer 구조가 높은성능
- 사전학습에서 noising된 input을 denoising하여 단어 예측하는 방식이 가장 효율적 (MLM)
- Domain specific data는 task에 도움되지만 data 크기 작으면 overfitting 야기
- Multitask learning이 비지도 학습과 비슷한 성능
- 작은 모델을 큰 데이터로 학습
- 파라미터 클수록 성능 좋음

3. Training obj. : Modified MLM

Masked

MLM은 bidirectional model 구조

BERT는 하나의 token에 masking하지만 T5는 연속 token을 하나의 mask로 (BART와 유사)

↳ original: Thank you ~~for inviting~~ me to your party ~~last~~ week.

input: Thank you <X> me to your party <Y> week.

target : ⟨X⟩ for inviting ⟨Y⟩ last ⟨Z⟩

input에서 masking 되지 않았던 부분을 맞히도록 (MASS 논문에서)

Output은 FFNN + Softmax 통해 sequence 생성