XLNet : Generalized Autoregressive Pretraining for Language Understanding

https://arxiv.org/pdf/1906.08237

0. Introduction

- 자연어처리(NLP)에서 사전학습 기반 표현 학습은 대규모 말뭉치를 활용해 downstream task 성능을 높이는 핵심 기술임.
- 기존 접근법은 Auto-Regressive(AR) 모델(GPT)과 Auto-Encoding(AE) 모델(BERT) 기반으로 나뉨.
- BERT는 양방향 문맥을 학습하지만, 입력 마스크와 사전학습-파인튜닝 간 불일치 문제, 마스크된 토큰 간 의존성 학습 한계 존재.
- 기여: XLNet은 generalized autoregressive pretraining 방법을 제안, AR과 AE의 장점을 결합하고 BERT의 한계를 극복함.

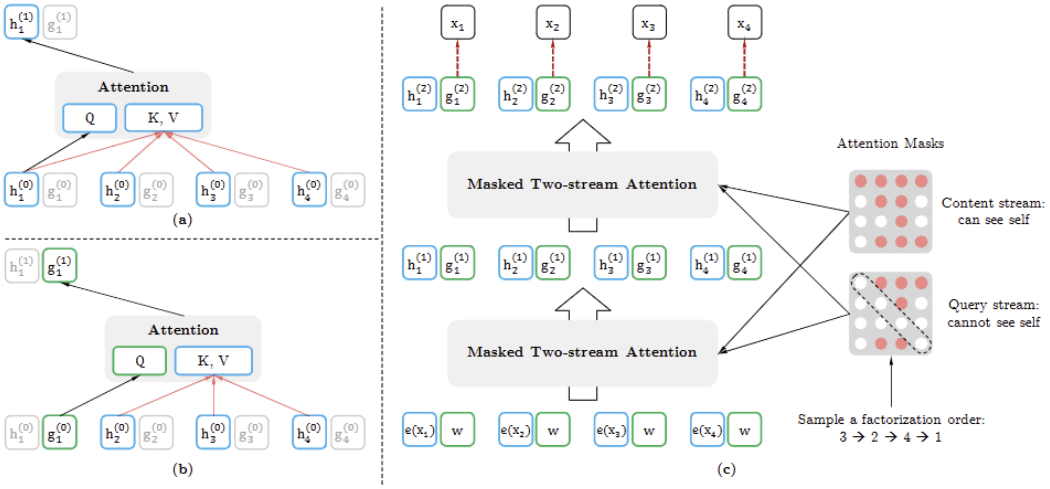
1. Overview

- XLNet은 Permutation Language Modeling 기반으로 모든 가능한 토큰 순서(permutation)에 대해 기대 우도를 최대화함으로써 양방향 문맥을 학습.
- 두 가지 핵심 요소:
  - Permutation-based training objective: 토큰 순서를 다양한 방식으로 재배열해 문맥 양쪽 모두 학습
  - Two-stream self-attention + Transformer-XL 기반 아키텍처: 긴 문맥 모델링과 지속적 컨텍스트 유지 가능
- XLNet은 사전학습 후 fine-tuning으로 다양한 NLP task에서 BERT보다 성능 우수.

2. Challenges

- BERT의 한계: 마스킹 기반 학습은 사전학습과 파인튜닝 간 불일치 문제를 야기, 마스크된 토큰 간 의존 관계 충분히 반영 못함.
- 전통적 AR 모델: 입력의 앞쪽 또는 뒤쪽만 문맥으로 사용하여 단방향 컨텍스트 한정.
- 양방향 문맥 학습 체계 설계 난제: 모든 토큰의 조합을 고려하는 과정에서 모호한 대상 예측 문제 발생.

3. Method



- Permutation Language Modeling Objective: 시퀀스의 모든 가능한 순열에 대한 우도 기대치를 최대화하여 양방향 문맥 정보 학습.
- Two-Stream Self-Attention: 각 토큰 위치에 대해 content stream과 query stream을 별도로 유지하여, 대상 토큰 콘텐츠와 문맥 기반 정보 분리 학습.
- Transformer-XL 통합: segment recurrence와 relative positional encoding을 사전학습에 적용해 긴 문맥 의존성 처리 능력 강화.
- 모델 구조: 기존 Transformer 기반 모델에 두 스트림과 permutation objective를 결합, AR과 AE 장점 활용.

4. Experiments

- 사전학습 데이터: BooksCorpus, English Wikipedia, Giga5, ClueWeb 2012-B, Common Crawl 등 대규모 텍스트 말뭉치.
- 비교 대상: BERT 및 기타 사전학습 기반 모델.
- 평가: GLUE, SQuAD, RACE, 감정 분석, 문서 순위 등 다양한 NLP 태스크 약 20개에서 성능 평가.
- 평가 지표: NLP 과제별 표준 지표(정확도, F1 점수 등).

5. Results

Model	SQuAD1.1	SQuAD2.0	RACE	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B
BERT-Large (Best of 3)	86.7/92.8	82.8/85.5	75.1	87.3	93.0	91.4	74.0	94.0	88.7	63.7	90.2
XLNet-Large-wikibooks	88.2/94.0	85.1/87.8	77.4	88.4	93.9	91.8	81.2	94.4	90.0	65.2	91.1

RACE	Accuracy	Middle	High	Model	NDCG@20	ERR@20
GPT [28]	59.0	62.9	57.4	DRMM [13]	24.3	13.8
BERT [25]	72.0	76.6	70.1	KNRM [8]	26.9	14.9
BERT+DCMN* [38]	74.1	79.5	71.8	Conv [8]	28.7	18.1
RoBERTa [21]	83.2	86.5	81.8	BERT†	30.53	18.67
XLNet	**85.4**	**88.6**	**84.0**	XLNet	**31.10**	**20.28**

SQuAD2.0	EM	F1	SQuAD1.1	EM	F1
Dev set results (single model)					
BERT [10]	78.98	81.77	BERT† [10]	84.1	90.9
RoBERTa [21]	86.5	89.4	RoBERTa [21]	88.9	94.6
XLNet	**87.9**	**90.6**	XLNet	**89.7**	**95.1**
Test set results on leaderboard (single model, as of Dec 14, 2019)					
BERT [10]	80.005	83.061	BERT [10]	85.083	91.835
RoBERTa [21]	86.820	89.795	BERT* [10]	87.433	93.294
XLNet	**87.926**	**90.689**	XLNet	**89.898‡**	**95.080‡**

Model	IMDB	Yelp-2	Yelp-5	DBpedia	AG	Amazon-2	Amazon-5
CNN [15]	-	2.90	32.39	0.84	6.57	3.79	36.24
DPCNN [15]	-	2.64	30.58	0.88	6.87	3.32	34.81
Mixed VAT [31, 23]	4.32	-	-	0.70	4.95	-	-
ULMFiT [14]	4.6	2.16	29.98	0.80	5.01	-	-
BERT [35]	4.51	1.89	29.32	0.64	-	2.63	34.17
XLNet	**3.20**	**1.37**	**27.05**	**0.60**	**4.45**	**2.11**	**31.67**

Model	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	WNLI
Single-task single models on dev									
BERT [2]	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-
RoBERTa [21]	90.2/90.2	94.7	92.2	**86.6**	96.4	**90.9**	68.0	92.4	-
XLNet	**90.8/90.8**	**94.9**	**92.3**	85.9	**97.0**	90.8	**69.0**	**92.5**	-
Multi-task ensembles on test (from leaderboard as of Oct 28, 2019)									
MT-DNN* [20]	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0
RoBERTa* [21]	90.8/90.2	98.9	90.2	88.2	96.7	92.3	67.8	92.2	89.0
XLNet*	**90.9/90.9**†	**99.0**†	**90.4**†	**88.5**	**97.1**†	**92.9**	**70.2**	**93.0**	**92.5**

- XLNet은 여러 NLP 벤치마크에서 BERT보다 뛰어난 성능 보임.
- GLUE, SQuAD, RACE 등 자연어 이해 task에서 상대적으로 큰 성능 개선 관찰.
- 결과는 XLNet의 양방향 맥락 캡처 능력과 autoregressive 목표가 downstream task에 잘 일반화됨을 시사.
- 모델 복잡성 증가에도 다양한 언어 과제에서 robust 성능 확인.

6. Insight

- Permutation 기반 AR pretraining은 양방향 문맥 정보를 효과적으로 캡처하면서 사전 학습-파인튜닝 불일치 문제 완화.
- Two-stream self-attention과 Transformer-XL 통합 설계는 긴 문맥 처리에 강점 제공.
- XLNet 접근 방식은 순방향/역방향 컨텍스트를 동시에 반영해 AR과 AE의 장점을 결합한 효과적 pretraining 전략.
- 향후 연구: 효율적인 permutation sampling 전략 및 모델 경량화 연구 필요(추론 비용과 학습 효율 개선 측면).