

Summary of Class

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

In this class,

- NMT를 통해 자연어 생성의 기초부터 고난도 내용까지 다루었음
 - 수식에 기초한 이론부터 경험에 기반한 인사이트를 아우르는 내용
 - 전처리부터 구현, 객관적인 평가 방법에 이르기까지 실제 업무/연구 과정을 그대로 다루었음
- 실제 프로젝트/연구를 진행하기 위한 구조 설계 및 코드 구현
 - 다양한 알고리즘/하이퍼파라미터를 실험 해보기 위한 테스트베드 구현

Language Modeling

Intuition and Equations

- 우리가 사용하는 언어의 확률 분포를 모델링
 - 문장 전체에 대한 확률
 - 이전 단어들이 주어졌을 때, 다음 단어의 확률

$$\begin{aligned}P(x_{1:n}) &= P(x_1, \dots, x_n) \\&= P(x_n | x_1, \dots, x_{n-1}) \cdots P(x_2 | x_1) P(x_1) \\&= \prod_{i=1}^n P(x_i | x_{<i})\end{aligned}$$

$$\log P(x_{1:n}) = \sum_{i=1}^N \log P(x_i | x_{<i})$$

n-gram

- Exact matching word sequence에 대해 counting하여 approximation
 - Markov Assumption을 도입하여 전체 대신 일부만 counting
 - 쉽고 강력하지만, unseen word sequence에 대한 generalization 부족
- Smoothing, Discounting을 통해 generalization 능력 보완
 - 근본적인 해결책이 아님
 - NNLM을 통해 해결 가능

Sequence to Sequence with Attention

구성 요소

- Encoder
 - 문장을 context vector로 압축
- Decoder
 - Context vector를 문장으로 변환
- Generator
 - Decoder의 hidden state를 확률 분포로 변환

Attention

- Key-value function
 - Key와 잘 매칭되도록 Query를 잘 변환하는 방법을 학습하는 것
- QKV in Seq2seq
 - 디코더의 현재 상태(query)에서 필요한 정보를
 - 인코더의 출력(key)에서 검색하여
 - 가중치(attention weight)에 따라, 인코더의 출력(value)을 가중 합을 취함

Transformer

Multi-head Attention

- 여러 QKV를 통해 다양한 정보를 취득
 - Seq2seq와 달리 time-step에 병렬 계산
- Self-Attention
 - 인코더/디코더의 이전 레이어에서 정보를 취합하는 방법
 - 디코더는 미래를 보는 것을 막기 위한 masking 필요
- Attention
 - 디코더에서 인코더의 정보를 얻기 위함
 - Src 문장의 빈 칸에 대한 masking 필요

Optimization

- 최적화가 까다로움
 - LR warm-up & linear decay를 통해 학습 초반의 불안정함을 극복하려 함
 - 이로 인해 많은 hyper-param 추가
 - 더욱이 hyper-param에 민감함
- 다양한 연구들이 시도됨
 - Rectified Adam
 - Pre-LN Transformer

NLG is Auto-regressive Task

Auto-regressive 속성

- Teacher forcing으로 인한 학습과 추론의 괴리 발생
 - Input-feeding 등을 통해 완화
- Beam search를 통해 greedy search의 단점 보완
 - 데이터의 추가, 모델 개선 없이 성능 향상

RL to enhance NLG

- BLEU를 통해 최적화 가능
 - PPL은 실제 번역 품질을 정확하게 반영하지 못함
 - 기존의 방식에선 BLEU는 미분을 할 수 없으므로 Gradient based opt 불가
- 샘플링 기반 방식을 통해 학습
 - 학습과 추론의 방식이 같아짐
- 비효율적인 학습
 - Reward는 방향이 없음
 - 샘플링에 의존적

Leverage Monolingual Corpus

Back Translation

- Parallel corpus는 한정된 자원
 - 이를 통해 학습된 LM은 풍부한 표현력이 부족함
- 반대쪽 모델을 활용하여 합성 코퍼스 생성
 - 합성 코퍼스를 통해 디코더가 풍부한 표현 학습 가능
- Tagged BT 등을 통해 기존 BT의 한계 극복 가능
 - Original corpus와 synthetic corpus의 비율

Dual Learning

- $x \rightarrow y, y \rightarrow x$ 두 관계를 leverage
 - BT를 수학적으로 해석 가능한 방법
 - 정보의 손실이 거의 없는 NMT에서 강점
 - DSL, DUL, Dual Learning
- RL에 비해 좀 더 효율적인 학습 가능
 - MLE framework 내에서 해결

Final Wrap-up

- 이 수업의 내용 대부분을 이해하고 따라올 수 있다면 큰 자부심을 가져도 된다고 생각합니다.
- 이 수업을 통해서 단순히 이론/개념/수식/구현 뿐만 아니라, 업무/연구를 진행하는 방법도 배울 수 있었기를 바랍니다.
- 앞으로 배워갈 DL/NLP를 깊게 이해하고 활용하는데 큰 도움이 되길 바랍니다.