Week3. NLP models 3-3. Why and how to read papers

신승진

Contents

이유 2. 논문 읽는 방법 소개

3. 논문 기록 방법 소개 - 과제 제출 방법 소개

1. 왜 읽는가?

“우리는 응용 과학자입니다.”

- NLP Researcher/Engineer == 응용 과학자
    - 새로 나온 기술을 가장 먼저 내가 풀고자 하는 문제를 풀기 위한 도구로 적용하는 사람
- 기술 전파 속도
    - 논문, 기업 기술 블로그, 트위터 >>>>> 학교 강의, 책, 강좌
- 프로세스
    - 새로운 모델의 등장 → 논문 읽기 + (오픈 소스 코드 확인) → 내가 가지고 있는 데이터셋 or 목적 함수에 맞게 변형 또는 적용

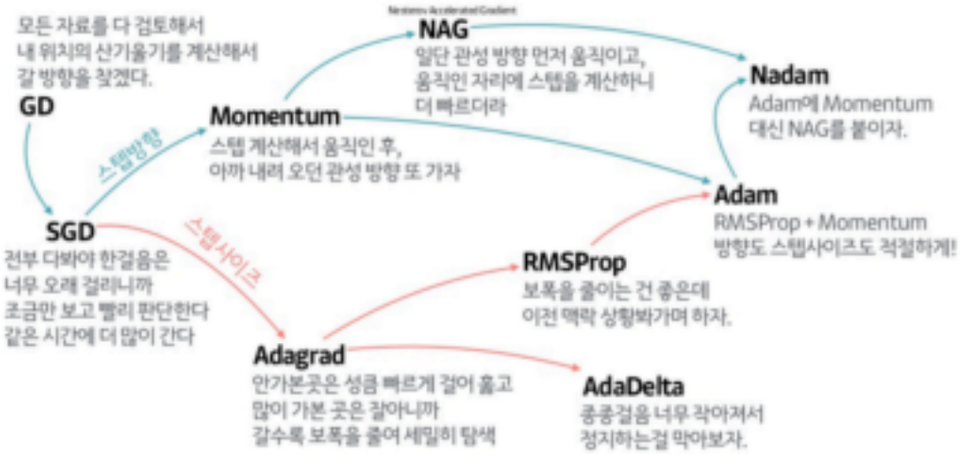
- 위 프로세스를 무한 반복

2. 어떻게 읽는가?

“미시적으로, 거시적으로”

- [Basic🐣] 이분법적으로 접근하기
- 문제점 ↔️ 해결점
- 과거 ↔️ 현재
- [Advanced🐥] 큰 흐름을 이해하고 해당 논문을 범주화 하기 • 방향성을 가지 치기

<Optimizer의 종류>



출처 : https://www.slideshare.net/yongho/ss-79607172

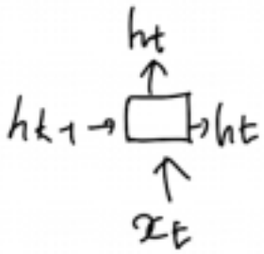
2. 어떻게 읽는가?

ex. “Attention is all you need”

2. 어떻게 읽는가?

ex. “Attention is all you need”

transduction problems such as language modeling and machine translation [35, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [38, 24, 15].



Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t, as a function of the previous hidden state h_{t-1} and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. Recent work has achieved significant improvements in computational efficiency through factorization tricks [21] and conditional computation [32], while also improving model performance in case of the latter. The fundamental constraint of sequential computation, however, remains.

Attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of dependencies without regard to their distance in the input or output sequences [2, 19]. In all but a few cases [27], however, such attention mechanisms are used in conjunction with a recurrent network.

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

Limit ①
Limit Ongoing ②
한계인 이유: RNN계열모델의 긴문장 Dependency 문제를 해결함 by Attention Only Model (~~RNN~~)
+. Parallel O
+. Quality ↑ (BLEU)
+. Training Time ↓ (12H 8GPUs)

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building

3. 어떻게

기록하는가?“머리 속으로

아는 것 ≠ 글로 정리하는 것” • Github

Repository

- Format: markdown file

- Github blog

- Velog

- History

4. Reference

- [blog on reading nlp papers](https://towardsdatascience.com/how-to-learn-deep-learning-by-reading-papers-c51b6025f226)
  ⭐⭐