급변하는 여론동향 파악을 위한 텍스트 분산 임베딩 방법

2018. 11. 30
한국BI데이터마이닝 학회 2018 추계학술대회 특별 세션
류성원, 김지영, 김누리, 이지훈, 조성준
lyusungwon@dm.snu.ac.kr
jeeyung.kim@dm.snu.ac.kr
noori@dm.snu.ac.kr
t080205@gmail.com
zoon@snu.ac.kr

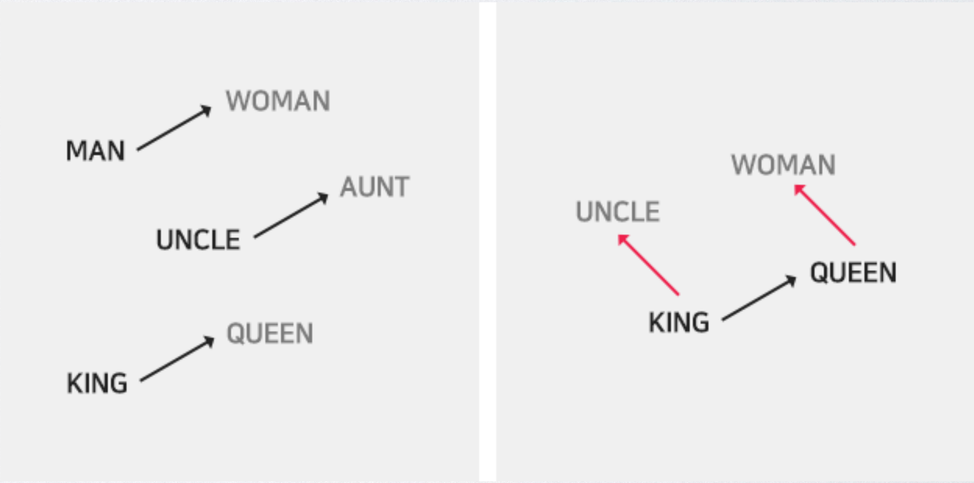
목차

서론

- 인터넷 상의 웹사이트 글, 뉴스 기사, 포털 댓글, SNS 등의 텍스트 데이터는 국민의 의견과 감정을 반영하기 때문에 인터넷의 텍스트를 분석하는 일은 국가 안보에 매우 중요함
- 여론 파악을 위한 인터넷의 텍스트 분석의 요구조건
  - 1) 많은 양의 텍스트들에 담긴 정보를 풍부하면서도 압축적으로 담아야 함
  - 2) 이에 대한 처리가 합리적으로 빨라야 함
  - 3) 다양한 출처의 데이터들을 실시간으로 반영할 수 있어야 함

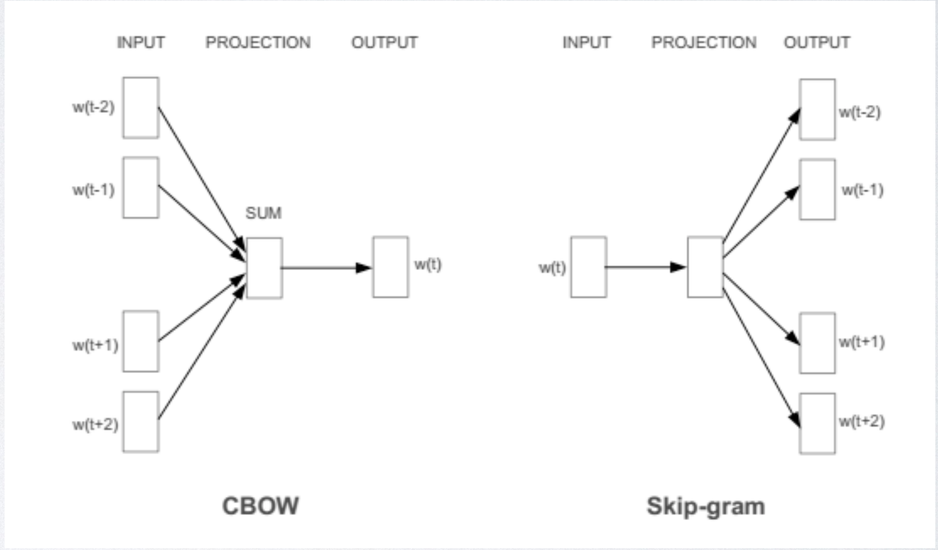
관련연구 - 단어 임베딩

- 기존 단어 임베딩
  - 단어 간의 동시 발생 빈도수를 통하여 단어를 표현
  - 단점: 이산적 표현은 분포가 희박하고 정교한 의미를 포함하기 어려움
- 신경망 기반 단어 임베딩
  - 알고리즘 기반 (Skipgram, CBOW, GloVE)
  - 연속적 표현
  - 의미론적 정보와 구문론적 정보를 반영



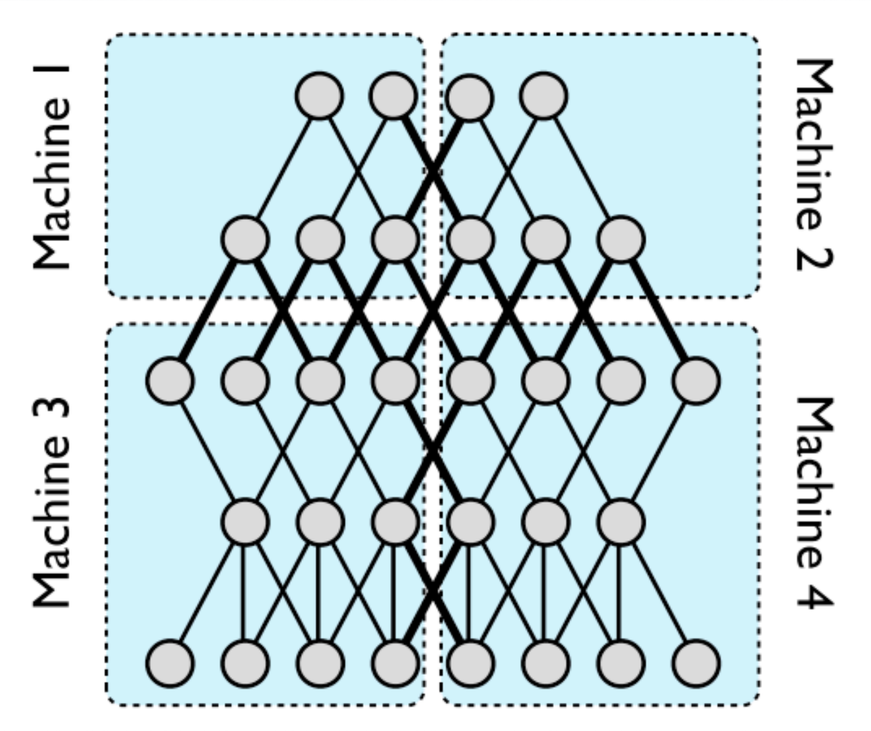
관련연구 - 단어 임베딩

- 임베딩 알고리즘
  - CBOW
    - 주변 단어로 중심 단어를 예측
  - Skip-Gram and Negative Sampling
    - 중심 단어로 주변 단어를 예측 & 부정 표본 추출 기법
  - GloVE
    - 발생 빈도수에 대한 정보 반영
  - FastText
    - N-gram 단위 임베딩

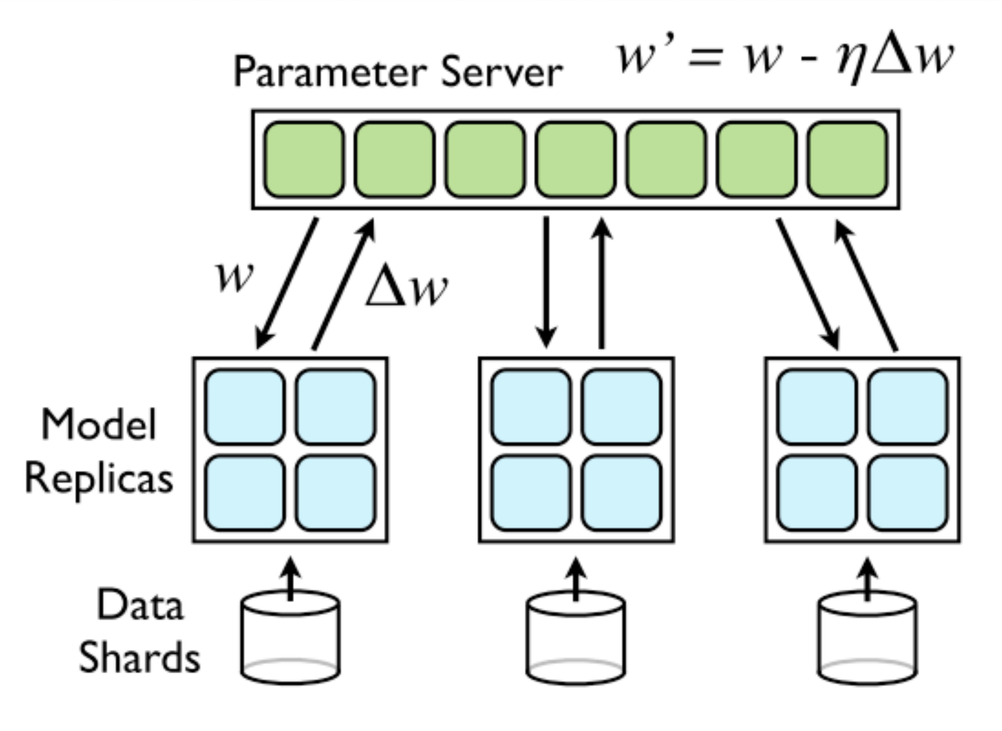


관련연구 - 분산 딥러닝 학습

- 딥러닝의 분산
  - 모델 분산

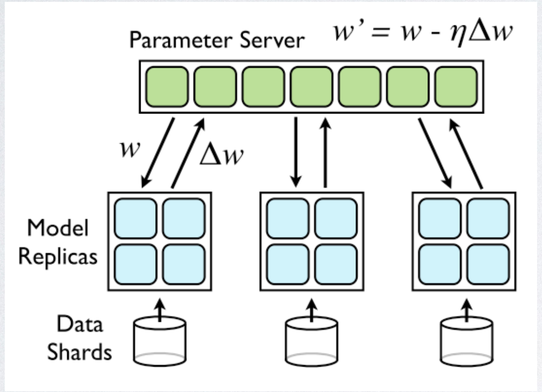


  - 데이터 분산



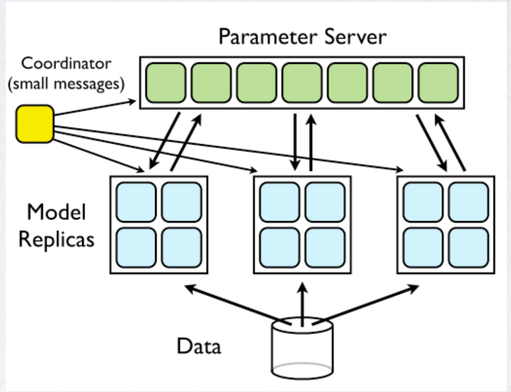
관련연구 - 분산 딥러닝 학습

- 데이터 분산 동기화 방식
  - 비동기적 동기화



    - 각 모델의 학습이 완료되는대로 업데이트
    - 시간당 데이터 처리속도가 빠름
    - 그래디언트를 구한 파라미터와 업데이트 하려는 모델의 파라미터 사이의 괴리가 발생

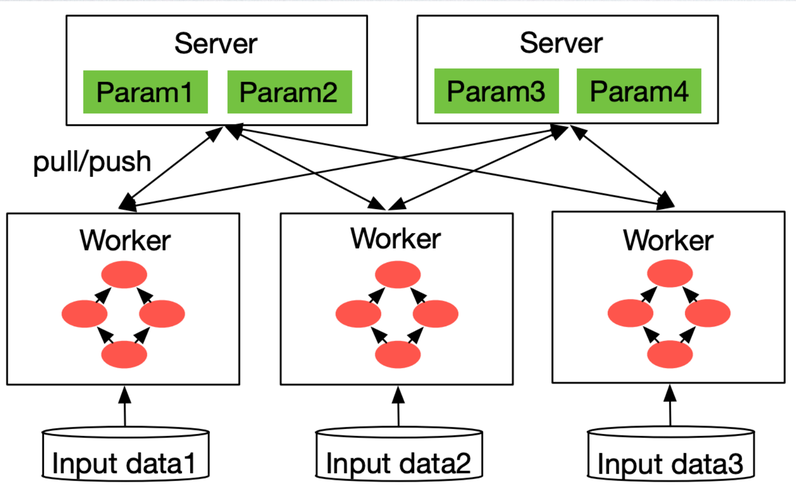
  - 동기적 동기화



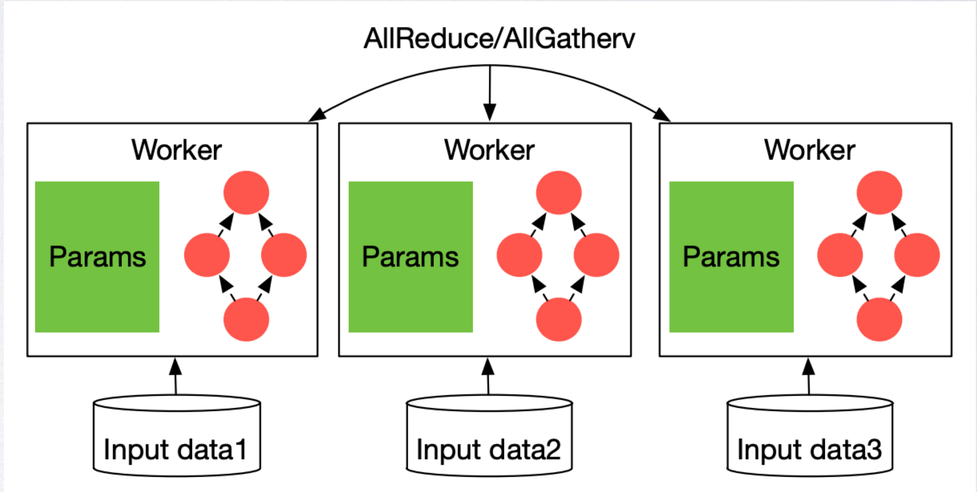
    - 모든 모델의 학습 후 함께 업데이트
    - 가장 느린 노드의 속도에 맞춰짐
    - 그래디언트를 정확하게 구하기 때문에 데이터 효율이 높음

관련연구 - 분산 딥러닝 학습

- 분산 학습 구조
  - 파라미터 서버 구조



    - 파라미터 서버와 워커 노드들을 분리
    - 워커 노드들은 서버와만 통신
    - 분포가 드문 그래디언트 업데이트에 유리
  - All-reduce, All-gather 구조



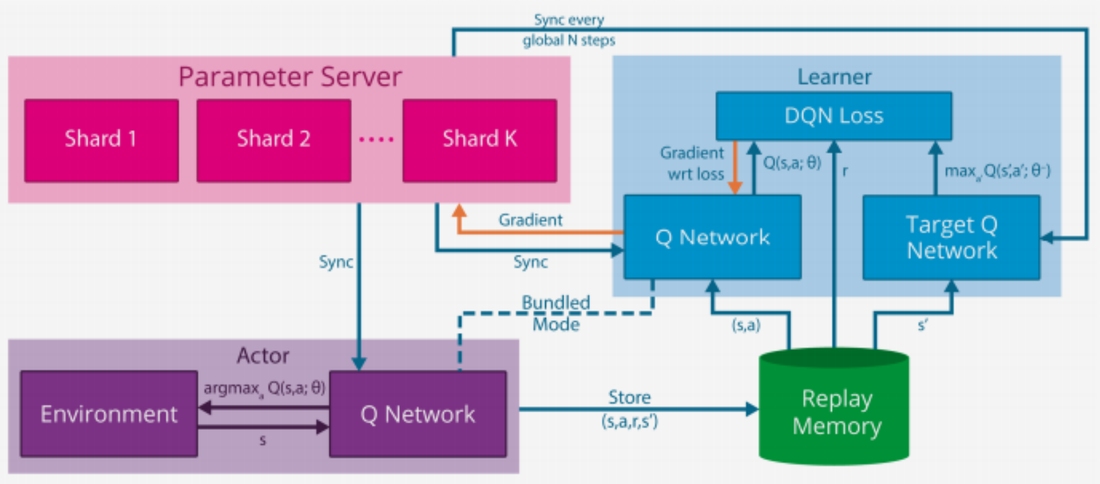
    - 파라미터 서버를 따로 두지 않음
    - 워커 노드들 끼리 통신
    - 분포가 조밀한 그래디언트 업데이트에 유리

관련연구 - 스트림 딥러닝 학습

- 기존 스트림 처리
  - 데이터의 생성과 처리 시간 사이의 간극을 처리
- 딥러닝의 특성
  - 모델 학습과 추론이 분리
  - 스트림 학습에 대한 연구가 희박
- 강화학습
  - 환경에서 얻은 경험을 통하여 모델을 학습하고 이를 통하여 환경에서 다른 경험을 얻음
  - 데이터들을 즉각적으로 반영해야 함
  - 분산 처리에 대한 연구가 많음

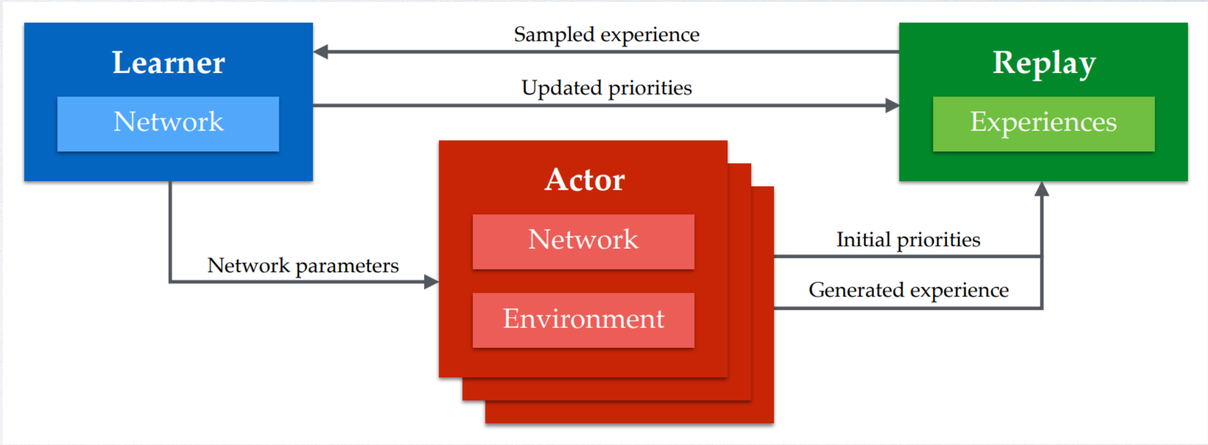
관련연구 - 스트림 딥러닝 학습

- 분산 강화학습
  - Gorila Framework



    - PS 분산 처리 구조
    - Actor, Learner, Parameter Server를 분리
    - 재생 기억 저장소를 통한 데이터 분배

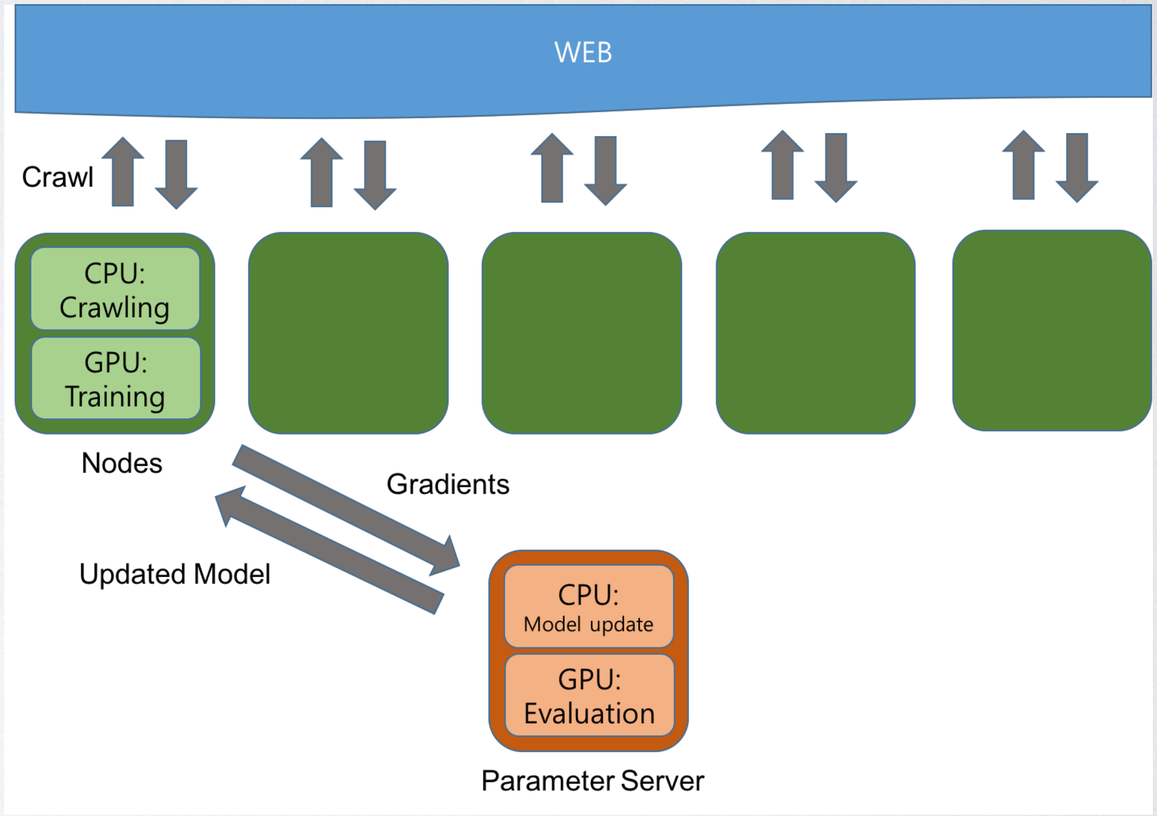
  - Ape-X Framework



    - PS를 따로 두지 않고 Actor만 분산
    - 학습은 Learner에서만 진행
    - 재생 기억 저장소에서 우선화 기반 표본 추출

제안 방법

- 텍스트 분산 임베딩 방법
  - 파라미터 서버 분산 구조
  - 워커 노드
    - 크롤링 / 스트리밍
    - 그래디언트 계산
  - 파라미터 서버
    - 모델 업데이트 및 평가
- 환경에 따른 동기화 방식 선택
- 데이터 표본 추출 방식 선택



예상 결과

- 기존 텍스트 임베딩의 문제점
  - 많은 양의 텍스트 데이터들이 필요
  - 텍스트 데이터 수집 시간이 오래 걸림
  - 큰 데이터 저장소가 필요
  - 학습 시간이 오래 걸림
  - 새로운 데이터를 반영할 수 없음
- 제안 방법
  - 데이터 수집 시간 단축
  - 디스크 저장소를 거치지 않기 때문에 큰 저장소가 필요하지 않음
  - 분산 학습을 통한 시간 단축
  - 여러 출처의 데이터를 반영하여 편향 극복
  - 새로운 데이터를 지속적으로 반영

미래 연구 방향

- 선결 조건
  - 지속적으로 크롤링 / 스트리밍 할 수 있는 출처가 필요
- 개선 방안
  - 여러 데이터 표본 추출 방식의 성능 비교
  - 동기적 동기화를 위한 통신 속도 개선
  - 임베딩을 객관적으로 평가할 만한 지표 개발

참고 논문

- [1] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
- [2] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.
- [3] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
- [4] Joulin, Armand, et al. "Bag of tricks for efficient text classification." arXiv preprint arXiv:1607.01759 (2016).
- [5] Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.
- [6] Levy, Omer, and Yoav Goldberg. "Linguistic regularities in sparse and explicit word representations." Proceedings of the eighteenth conference on computational natural language learning. 2014.
- [7] Levy, Omer, and Yoav Goldberg. "Neural word embedding as implicit matrix factorization." Advances in neural information processing systems. 2014.
- [8] Levy, Omer, Yoav Goldberg, and Ido Dagan. "Improving distributional similarity with lessons learned from word embeddings." Transactions of the Association for Computational Linguistics 3 (2015): 211-225.

참고 논문

- [9] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.
- [10] Isard, Michael, et al. "Dryad: distributed data-parallel programs from sequential building blocks." ACM SIGOPS operating systems review. Vol. 41. No. 3. ACM, 2007.
- [11] Zaharia, Matei, et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012.
- [12] Toshniwal, Ankit, et al. "Storm@ twitter." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.
- [13] Zaharia, Matei, et al. "Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters." HotCloud 12 (2012): 10-10.
- [14] Dean, Jeffrey, et al. "Large scale distributed deep networks." Advances in neural information processing systems. 2012.
- [15] Recht, Benjamin, et al. "Hogwild: A lock-free approach to parallelizing stochastic gradient descent." Advances in neural information processing systems. 2011.
- [16] Chen, Jianmin, et al. "Revisiting distributed synchronous SGD." arXiv preprint arXiv:1604.00981 (2016).

참고 논문

- [17] Ho, Qirong, et al. "More effective distributed ml via a stale synchronous parallel parameter server." Advances in neural information processing systems. 2013.
- [18] Strom, Nikko. "Scalable distributed DNN training using commodity GPU cloud computing." Sixteenth Annual Conference of the International Speech Communication Association. 2015.
- [19] Parallax: Automatic Data-Parallel Training of Deep Neural Networks. Soojeong Kim, Gyeong-In Yu, Hojin Park, Sungwoo Cho, Eunji Jeong, Hyeonmin Ha, Sanha Lee, Joo Seong Jeong, Byung-Gon Chun. arXiv:1808.02621, August 2018.
- [20] Nair, Arun, et al. "Massively parallel methods for deep reinforcement learning." arXiv preprint arXiv:1507.04296 (2015).
- [21] Mnih, Volodymyr, et al. "Asynchronous methods for deep reinforcement learning." International conference on machine learning. 2016.
- [22] Alain, Guillaume, et al. "Variance reduction in SGD by distributed importance sampling." arXiv preprint arXiv: 1511.06481 (2015).
- [23] Schaul, Tom, et al. "Prioritized experience replay." arXiv preprint arXiv:1511.05952 (2015).
- [24] Horgan, Dan, et al. "Distributed prioritized experience replay." arXiv preprint arXiv:1803.00933 (2018).

EOD