Laboratory Notebook

Research Note - Word2Vec Implementation Project

Page 1

Continued from page: N/A

관련 프로젝트: Word2Vec 재구현 프로젝트

실험 내용: Efficient Estimation of Word Representations in Vector Space 논문 기반 Word2Vec 모델 구현 및 학습

2024년 9월 첫째주

일자: 2024년 9월 첫째주

실험 목적:

- Efficient Estimation of Word Representations in Vector Space 논문 읽고 정리
- 서버 환경 세팅
- CBOW와 Skip-gram의 기본적인 기능과 구조 파악

실험 절차 및 결과:

1. **논문 학습**
    - CBOW와 Skip-gram의 기본적인 기능과 구조를 파악
    - 이전 연구들과의 차별점 및 보완된 부분 학습
    - 논문 읽기가 어색하여 예상보다 시간이 많이 소요됨
2. **서버 환경 구축**
    - Docker 환경 구성 시도
    - 서버의 PID 1 프로세스가 systemd가 아닌 sshd로 동작하여 Docker 데몬 실행 불가
    - 대안으로 conda로 가상환경 생성하여 진행

관찰 사항:

- 논문을 읽는 것이 어색하여 예상보다 시간이 많이 소요됨
- 이를 보완하기 위해 딥 러닝을 이용한 자연어 처리 입문 위키독스와 논문을 병행하며 Word2Vec 학습 계획

다음 주 계획:

- 위키독스 Word2Vec 파트 읽기
- PyTorch 강의 수강
- CBOW 코드 재구현
- Skip-gram 코드 재구현

- 밑바닥부터 시작하는 딥러닝 Ch3-6 학습

2024년 9월 둘째주

일자: 2024년 9월 둘째주

실험 목적:

- 위키독스 Word2Vec 이론 학습 및 실습
- 딥러닝 기초 학습

실험 절차 및 결과:

1. **위키독스 학습**
    - 09-01~09-06 학습 및 실습 진행
    - Word2Vec 이론 학습과 여러 구현된 모델들을 실습
    - 모델이 작동하는 방식을 이해하는데 도움이 됨
    - 직접 모델들을 구현한 뒤 평가할 때 좋은 지표들이 될 것으로 예상
2. **딥러닝 기초 학습**
    - 밑바닥부터 시작하는 딥러닝 Ch3 학습

관찰 사항:

- 위키독스 내 Word2Vec 이론 학습과 여러 구현된 모델들을 실습하는 과정이 생각보다 오래 걸림
- 점심시간 직후엔 학습에 집중이 잘 되지 않아서 여러 대안들을 시도 중

다음 주 계획:

- 위키독스 09-08~09-14 학습 및 실습
- PyTorch 강의 수강
- CBOW 코드 재구현
- Skip-gram 코드 재구현

2024년 9월 셋째주

일자: 2024년 9월 셋째주

실험 목적:

- CBOW 클론 코딩
- Skip-gram 클론 코딩
- 딥러닝 기초 학습

실험 절차 및 결과:

1. **CBOW 클론 코딩**
    - PyTorch로 구현되어 있는 코드들을 그대로 구현

    - 스스로 어떻게 구현해야 할지 감을 익힘
    - 여러 값들을 바꿔보며 모델로 구현된 gensim과 비교

2. **Skip-gram 클론 코딩**

    - PyTorch로 구현되어 있는 코드들을 그대로 구현

3. **딥러닝 기초 학습**

    - 밑바닥부터 시작하는 딥러닝 Ch4 학습

관찰 사항:

- 처음 구현해볼 때 학습 시간이 유난히 오래 걸려서 gpu 사용량을 확인
- device를 cuda로 설정하지 않아서 cpu를 사용하고 있었음
- 이런 시행착오를 겪으면서 코드 재구현이나 교재 학습이 예상보다 오래 걸리는 것을 알게 됨
- 점심식사 이후의 시간에 대해선 출력값 비교처럼 집중력을 요구하지 않는 작업들을 하며 루틴을 개선 중

다음 주 계획:

- CBOW 직접 구현
- Skip-gram 직접 구현
- 밑바닥부터 시작하는 딥러닝 5, 6 학습
- Word2Vec 프로젝트 정리 및 paper 작성

2024년 9월 넷째주

일자: 2024년 9월 넷째주

실험 목적:

- CBOW 직접 구현
- Skip-gram 직접 구현
- 언어에이전트 발표 준비

실험 절차 및 결과:

1. **CBOW 직접 구현**

    - PyTorch로 직접 구현하는 것에 집중하다가 Instruction에 적힌 요구사항을 잊고 있던 것을 발견
    - 디렉토리 구조에 맞춰 코드를 나누고 테스트
    - 랜덤시드를 고정하고, configs 내 yaml 파일로 실행하도록 설정
    - 해당 과정을 진행하면서 ssh환경에 github를 연결 (key를 생성하고 등록)

2. **Skip-gram 직접 구현**

    - 진행 중

3. **언어에이전트 발표 준비**

    - 논문을 읽고 지식을 습득하는 게 익숙하지 않아서 상당히 시간이 소요됨
    - 새로운 avsr에 대해 알아가는 과정

관찰 사항:

- 디렉토리 구조에 맞춰 코드를 나누고 테스트하느라 CBOW만 작업하는 데 시간이 꽤 소요됨
- 랜덤시드를 고정하고, configs 내 yaml 파일로 실행하도록 설정하는 게 처음이라 많이 배움
- ssh환경에 github를 연결하는 과정이 조금 번거로웠지만 덕분에 작업하기 수월해짐

다음 주 계획:

- Skip-gram 직접 구현 마무리
- Word2Vec 프로젝트 정리 및 paper 작성
- 언어에이전트 발표

2024년 10월 셋째주

일자: 2024년 10월 셋째주

실험 목적:

- Word2Vec 재구현 및 평가
- 언어에이전트 Trolley Problem 논문 재구현 발표 준비

실험 절차 및 결과:

1. **Word2Vec 재구현 및 평가**
    - Word2Vec 구현한 뒤 테스트하는데 결과가 자꾸 이상하게 나와서 수정하느라 시간이 오래 걸림
2. **언어에이전트 발표 준비**
    - 논문 재구현하는 과정이 Word2Vec 구현에도 도움이 됨
    - 논문에서 주장한 연구 질문들에 대한 결과를 직접 받아보고, 실험 결과를 국가로 분류하여 해석

관찰 사항:

- Word2Vec 구현이 이렇게 오래 걸릴줄 몰랐음
- 언어에이전트 논문 재구현하는 과정이 Word2Vec 구현에도 도움이 됨

다음 주 계획:

- Word2Vec 프로젝트 정리 및 paper 작성
- 언어에이전트 발표 준비
- ELMo 개념 학습

2024년 10월 넷째주

일자: 2024년 10월 넷째주

실험 목적:

- Word2Vec 재구현 및 평가
- Word2Vec 대규모 코퍼스 시도

실험 절차 및 결과:

1. **대규모 코퍼스 학습 시도**
    - 데이터셋 크기가 2GB 정도로 작아서 대규모 코퍼스로 학습한 논문의 흐름과 맞게 8GB정도로 크기를 키움
    - 코퍼스 규모가 커지니 코퍼스를 메모리에 올려두고 로드해서 학습하던 이전 방식으로 진행할 수 없음
    - 스트리밍 방식, mmap 등으로 코드를 수정
    - 대규모 코퍼스를 안정적으로 학습하는 코드가 완성되지 않아서 학습이 제대로 되지 않는 상태

관찰 사항:

- 대규모 코퍼스 학습을 진행하는 경험을 해보고 싶어서 계속 시도를 해봤지만, 이젠 결과를 내고 분석하는 과정도 중요한 것 같다는 판단
- 우선 크기가 작은 코퍼스(3GB)로 학습하면서 요구사항에 맞는 실험을 진행하고 결과를 분석하려 함
- Word2Vec 프로젝트를 마친 뒤 ELMo 개념을 학습할 때 대규모 코퍼스로 1epoch정도만 돌려볼 경험을 해보려 함

다음 주 계획:

- Word2Vec 실험 및 마무리
- ELMo 개념 학습
- 밑바닥부터 시작하는 딥러닝 ch5 학습 및 실습

2024년 10월 다섯째주

일자: 2024년 10월 다섯째주

실험 목적:

- Word2Vec 스트리밍 방식 완성
- ELMo 개념 학습

실험 절차 및 결과:

1. **Word2Vec 스트리밍 방식 완성**
    - 소규모 코퍼스를 먼저 학습해서 실험하고 결과를 빨리 받으려고 했지만, 코퍼스가 1GB정도로 작지 않은 이상 스트리밍 방식을 사용해야 함
    - 코드 구현하고 대규모 코퍼스로 진행
    - Loss가 증가하는 현상이 생겨서 학습률을 낮추고 토큰 처리를 추가해서 해결
    - 1epoch 학습하는데 거의 24시간이 걸림
2. **ELMo 개념 학습**
    - Word2Vec 코드 구현이 이제서야 다 완성되어서 ELMo 개념을 학습 중

관찰 사항:

- 1epoch 학습하는데 거의 24시간이 걸려서 모든 세팅의 결과를 받으려면 시간이 걸릴 것 같음
- 실험 및 결과 분석 마치는대로 paper 제출 예정

- 이번주 목요일부터 대학원 원서접수 기간이라서 연구계획서를 작성해야 함
- 아직 분야를 정확하게 정한 건 아니지만 이참에 한번 찾아보려 함

다음 주 계획:

- ELMo 개념 학습
- Word2Vec 실험 돌리기
- 연구 분야 탐색

2024년 11월 첫째주

일자: 2024년 11월 첫째주

실험 목적:

- Word2Vec 실험
- ELMo 개념 학습
- 연구계획서 작성

실험 절차 및 결과:

1. **Word2Vec 실험**
    - 1epoch 학습하는데 약 10시간 정도 걸림
    - 학습 결과를 csv파일로 저장하는데, 파일 이름을 간단하게 저장하도록 설정했더니, seed 구분 없이 저장돼서 파일이 섞여서 애를 먹음
    - 다음에는 파일 저장 경로를 처음부터 명확하게 저장하도록 하려고 함
2. **ELMo 개념 학습**
    - 이번주부터는 ELMo 코드 구현도 시작하려고 함
3. **연구계획서 작성**
    - 입학 후에 문화와 맥락을 반영한 번역에 대해 연구해보려 함
    - 아직 이 분야에 대한 지식은 많지 않지만, 연구 계획서 작성을 위해 관련 논문들을 읽어보고 있음
    - 이번주 내로 대학원 원서 접수는 마침
    - 원서 접수 마친 뒤에도 분야 탐색은 계속 해보려 함

관찰 사항:

- 1epoch 학습하는데 약 10시간 정도 걸려서 시간이 좀 걸림
- 파일 저장 경로를 처음부터 명확하게 저장하도록 설정하는 것이 중요함

다음 주 계획:

- Word2Vec paper 제출
- ELMo 코드 구현
- 연구 분야 탐색

MEMO

메모는 이렇게 상하의 구분을 짓고, 자유롭게 무엇이든 작성할 수 있습니다!!

제 3자가 이해할 수 있는 과정이나 아이디어, 모든 과정에서 나왔던 것을 기록해보세요 ^_^

그림도 가능! 계산도 가능!

주요 학습 내용 및 인사이트

- **CBOW vs Skip-gram**: CBOW는 주변 단어로 중심 단어를 예측하는 모델이고, Skip-gram은 중심 단어로 주변 단어를 예측하는 모델임. Skip-gram이 희소한 단어에 대해 더 좋은 성능을 보임.
- **Negative Sampling**: 기존의 Hierarchical Softmax보다 계산 효율이 좋고 성능도 우수함. 샘플링 기법을 통해 학습 속도를 크게 향상시킬 수 있음.
- **대규모 코퍼스 처리**: 메모리에 모든 데이터를 올리는 방식은 한계가 있음. 스트리밍 방식이나 mmap을 활용하여 대규모 데이터를 효율적으로 처리해야 함.
- **학습 시간 최적화**: GPU 사용 여부를 항상 확인해야 함. device를 cuda로 설정하지 않으면 CPU를 사용하게 되어 학습 시간이 크게 증가함.
- **파일 관리**: 실험 결과를 저장할 때 seed, 하이퍼파라미터 등을 파일명에 명확하게 포함시켜야 나중에 결과를 분석할 때 혼란이 없음.

시행착오 및 개선 사항

1. **Docker 환경 구축 실패**: 서버의 PID 1 프로세스가 systemd가 아닌 sshd로 동작하여 Docker 데몬 실행 불가 → conda 가상환경으로 대체
2. **GPU 미사용 문제**: device를 cuda로 설정하지 않아 CPU로 학습하여 시간이 오래 걸림 → 항상 device 확인하는 습관 필요
3. **파일 저장 경로 문제**: 파일 이름을 간단하게 저장하여 seed 구분이 안 됨 → 파일 저장 경로를 처음부터 명확하게 설정
4. **Loss 증가 현상**: 대규모 코퍼스 학습 시 Loss가 증가하는 현상 발생 → 학습률을 낮추고 토큰 처리 추가하여 해결

향후 연구 방향

- 문화와 맥락을 반영한 번역 연구
- ELMo, BERT 등 최신 언어 모델 학습
- 대규모 코퍼스 학습 경험 축적

기록자 (Invented by): 양혜인

일자 (Date): 2024년 11월 첫째주

점검자 (Witnessed and Understood by): _____

일자 (Date): _____