

Report

김준호

word2vec.py

Word2Vec은 단어 임베딩을 학습하는 데 사용되며, CBOW(Continuous Bag of Words)와 Skip-gram 두 가지 방법을 지원한다.

Word2Vec 클래스

`__init__` 메서드:

- vocab_size**: 어휘 크기. 사용할 수 있는 고유 단어의 수이다.
- d_model**: 임베딩 차원. 각 단어를 표현할 벡터의 길이를 나타낸다.
- window_size**: CBOW 및 Skip-gram에서 문맥을 고려할 단어의 수를 정의한다.
- method**: "cbow" 또는 "skipgram" 중 하나를 선택하여 학습 방법을 지정한다.
- embeddings**: 단어를 벡터로 변환하는 임베딩 레이어이다.
- output_weights**: 임베딩 벡터를 다시 어휘 크기로 변환하는 선형 레이어이다.

embeddings_weight 메서드:

- 학습된 임베딩의 가중치를 반환한다. 모델이 학습한 단어 임베딩을 외부에서 사용할 수 있도록 역할을 한다..

fit 메서드:

- load_corpus를 기반으로 Word2Vec 모델을 학습한다.
- tokenizer를 사용하여 텍스트 데이터를 토큰화한다.
- CBOW와 Skip-gram 방법에 따라 각각 문맥 단어와 타겟 단어를 정의하고 학습을 진행한다.
- 각 에폭 동안 손실을 계산하고 이를 출력하여 학습 과정을 모니터링한다.

`_train_cbow` 메서드:

- CBOW 방식을 사용하여 학습한다.
- 문맥 단어의 평균 임베딩을 통해 타겟 단어를 예측한다.
- 손실을 역전파하여 가중치를 업데이트한다.

`_train_skipgram` 메서드:

- Skip-gram 방식을 사용하여 학습한다.
- 타겟 단어의 임베딩을 사용하여 문맥 단어를 예측한다.
- 손실을 역전파하여 가중치를 업데이트한다.

`load_corpus.py`

Word2Vec 모델을 학습시키기 위한 코퍼스를 제공한다. 간단한 예제 문장 리스트를 사용했다. 외부 소스로부터 데이터를 가져온다면 좀 더 제대로 할 수 있겠지만 시간이 너무 오래걸려 포기했다....

`load_corpus` 함수:

- 코퍼스를 문자열 목록으로 반환한다. 임의의 데이터셋을 사용할 수 있도록 설계되어 있다.

`gru.py`

Gated Recurrent Unit(GRU) 셀과 GRU 네트워크를 PyTorch로 구현한 것이다. GRU는 시퀀스 데이터를 처리하는 데 사용되며 LSTM보다 간단한 구조를 가지고 있다고 한다.

`GRUCell` 클래스

`__init__` 메서드:

-**input_size**: input의 차원이다.

-**hidden_size**: GRU의 hidden state 크기이다.

-**W_z, U_z**: update gate의 가중치이다.

-**W_r, U_r**: reset gate의 가중치이다.

-**W_h, U_h**: candidate hidden state의 가중치이다.

forward 메서드:

-입력 x 와 이전 hidden state h 를 사용하여 현재 hidden state h_t 를 계산한다.

-**업데이트 게이트 z_t** : z_t 입력과 이전 은닉 상태를 사용하여 게이트 값을 계산한다.

-**리셋 게이트 r_t** : r_t 입력과 이전 은닉 상태를 사용하여 게이트 값을 계산한다.

-**후보 은닉 상태 $(h_t)_{\tilde{}}$** : 리셋 게이트와 이전 은닉 상태를 활용하여 새로운 후보 은닉 상태를 계산한다.

-**새로운 은닉 상태 h_t** : 업데이트 게이트를 사용하여 이전 은닉 상태와 후보 은닉 상태를 결합하여 계산한다.

GRU 클래스

__init__ 메서드:

-**input_size**: 입력 특징의 차원이다.

-**hidden_size**: GRU의 은닉 상태 크기이다.

-**cell**: GRUCell 객체로 각 타임 스텝의 은닉 상태를 업데이트한다.

forward 메서드:

-입력 시퀀스를 순회하며 각 타임 스텝에서 GRU 셀을 통해 은닉 상태를 업데이트한다.

-마지막 타임 스텝의 은닉 상태를 반환한다. 이 은닉 상태는 시퀀스의 정보를 집약하여 담고 있다.