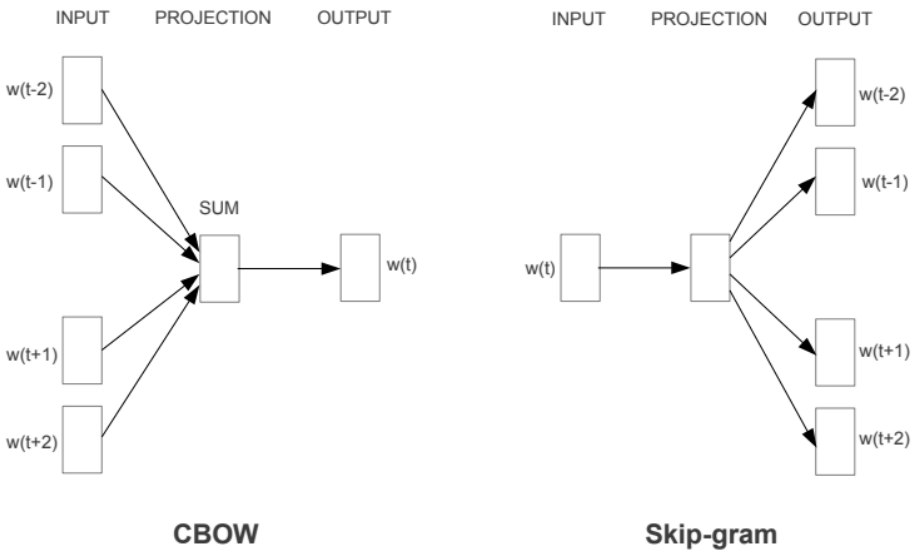
Word2Vec

Efficient Estimation of Word Representations in Vector Space

▼ Introduction

기존 자연어 처리 모델은 단어를 개별적인 단위로 취급하여 단어 간 유사성을 고려하지 않았다. 하지만 최근 연구에 따르면 분산 표현(word embeddings)을 활용하면 NLP 성능이 크게 향상된다. 본 논문에서는 대규모 데이터셋에서도 효율적으로 단어 벡터를 학습할 수 있는 두 가지 새로운 모델을 제안하며, 이를 통해 단어 간의 구문적(syntactic) 및 의미적(semantic) 관계를 효과적으로 보존할 수 있음을 보여준다.

▼ Proposed method



1. Continuous Bag-of-Words (CBOW) Model
    - 주변 단어들을 기반으로 특정 단어를 예측하는 방식
    - 단어 순서를 고려하지 않지만, 빠르고 효율적인 학습 가능
2. Continuous Skip-gram Model
    - 특정 단어를 입력으로 사용하여 주변 단어를 예측하는 방식

- 계산 비용이 더 크지만, 단어 간 의미적 관계를 더 정밀하게 학습할 수 있음

➡️ 두 모델 모두 전통적인 NNLM이나 LSA, LDA보다 뛰어난 성능을 보였다.

▼ Results & Performance

- 16억 개 단어로 이루어진 대규모 데이터셋을 하루 이내에 학습할 수 있을 정도로 높은 계산 효율성을 보임
- Skip-gram 모델이 특히 의미적 관계 학습에서 뛰어난 성능을 발휘함
  - 예: "King - Man + Woman ≈ Queen"과 같은 관계를 벡터 연산으로 유추 가능
- 학습된 단어 벡터는 구문 및 의미 유사성 테스트에서 SOTA을 기록함

➡️ 제안된 모델들은 대규모 자연어 처리(NLP) 작업에서 빠르고 확장 가능한 단어 임베딩 학습을 가능하게 한다.