GloVe(전역 벡터)는 "Global Vectors for Word Representation"의 약자로, 단어 임베딩을 생성하는 방법론입니다. GloVe는 단어의 의미를 벡터 공간 상에서 표현하는 기법으로, 단어 간의 의미 유사성을 수치적으로 표현할 수 있습니다.

GloVe는 크게 두 가지 개념에 기반하여 단어 벡터를 학습합니다: 동시 등장 행렬(co-occurrence matrix)과 임베딩 공간의 선형성(linearity of the embedding space).

동시 등장 행렬: GloVe는 말뭉치(corpus)에서 단어들이 함께 등장하는 빈도를 계산하여 동시 등장 행렬을 생성합니다. 동시 등장 행렬은 각 단어 쌍에 대해 해당 단어들이 함께 등장하는 빈도를 나타내는 행렬입니다. 이 행렬은 단어들 간의 의미적 관계를 잡아낼 수 있습니다.

임베딩 공간의 선형성: GloVe는 임베딩 공간 상에서 단어 벡터들의 선형 관계를 학습합니다. 예를 들어, "king - man + woman" 연산을 수행하면 "queen"에 가까운 벡터가 나오도록 학습됩니다. 이는 단어 간의 성질이 벡터 공간 상에서 선형적으로 표현될 수 있다는 가정에 기반합니다.

GloVe의 학습 과정은 동시 등장 행렬과 임베딩 벡터 간의 손실 함수를 최소화하는 방식으로 이루어집니다. 이 손실 함수는 단어 쌍의 동시 등장 횟수와 임베딩 벡터들 간의 내적 관계를 비교하여 계산됩니다. 학습이 완료되면, 각 단어는 고차원 벡터로 표현됩니다. 이 벡터들은 단어들의 의미적 유사성을 반영하고, 벡터 간의 유사성을 계산하여 단어 간의 관계를 파악하는 데 사용할 수 있습니다.

GloVe는 대규모 말뭉치에서 효과적으로 작동하며, 다양한 자연어 처리 태스크에서 좋은 성능을 발휘합니다. 또한, 미리 학습된 GloVe 임베딩을 사용하여 특정 태스크에 적용할 수도 있습니다. 이를 통해 단어의 의미를 벡터로 표현하고, 이를 활용하여 텍스트 데이터를 효과적으로 다룰 수 있습니다.