- UMAP (Uniform Maniford Approximation and Projection)

: 데이터의 복잡한 구조와 패턴을 유지하면서 차원을 축소하고 시각화하는 데 사용된다.

특징으로는 총 4가지로 정리할 수 있다.

1. Manifold Learning(매니폴드 학습) : 매니폴드란 고차원 공간에서 데이터가 분포하는 저차원 부분공간을 의미한다. 따라서 데이터의 복잡한 구조를 매니폴드로 추정하고 보존한다.
2. Local and Global Preservation(지역, 전역 보존) : 이웃 데이터 간의 거리 관계를 최대한 보존하면서도 전체적인 데이터 구조를 유지하려고 한다.
3. Randomized Approach(랜덤 기법 활용) : 확률적인 접근 방식을 사용하여 효율적인 방식으로 매니폴드를 학습하고 데이터를 저차원으로 투영시킨다.
4. 클러스터링 및 시각화 활용 : 클러스터링 및 데이터 시각화에 많이 사용되어, 고차원 데이터를 저차원으로 축소하면서 군집 구조를 보존하기 때문에, 데이터의 패턴을 파악하고 시각적으로 파악하는 데에 도움이 된다.

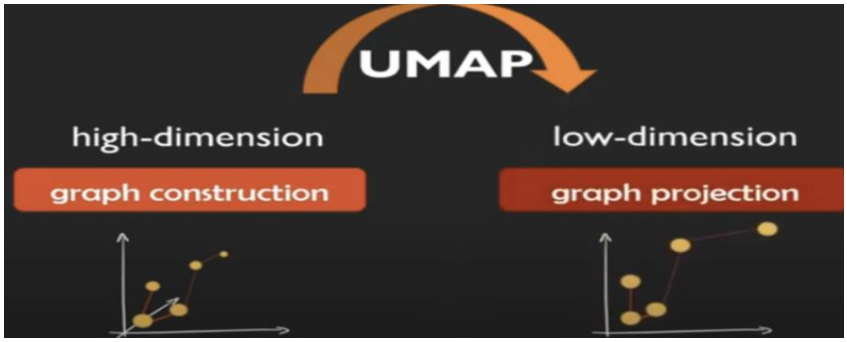
장/단점을 간략하게 짚고 넘어가면, 장점으로는 embedding 차원 크기에 대한 제한이 없어서 일반적인 차원 축소 알고리즘으로 사용하기 좋다. 또한 UMAP은 데이터 시각화뿐만 아니라 데이터 압축, 특성 선택 등 다양하게 활용할 수 있다. 단점으로는 Hyperparameter 값의 영향을 크게 받기 때문에 적절한 값을 선정하는 것이 중요하다. 또한 저차원 임베딩 시 정보손실에 의한 데이터 왜곡이 일어날 수 있다.

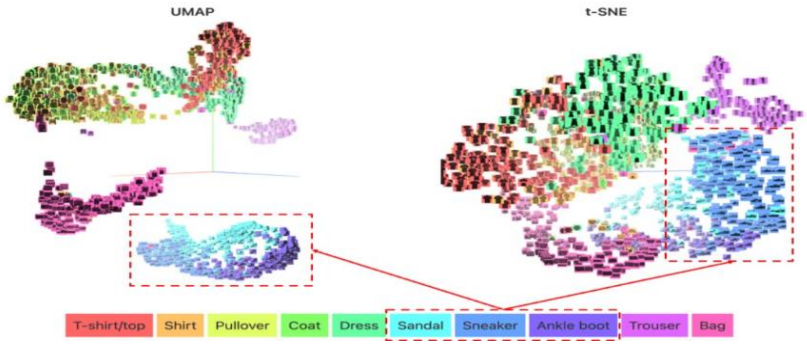
- PaCMAP(Pairwise Controlled Manifold Approximation and Projection)

: 데이터의 차원을 축소하고 시각화하기 위한 비지도 학습 기법 중 하나이다. 또한 가장 핵심적인 부분은 local / global str.를 유지하기 위해 차원 축소 방법론이 어떻게 작용하는지 확인하는 것이 중요하다. 여기서 local str. 유지를 위한 loss function 개념을 이해할 필요가 있다. 이는 데이터 간의 거리 관계를 나타내는 방식 중 하나이다.

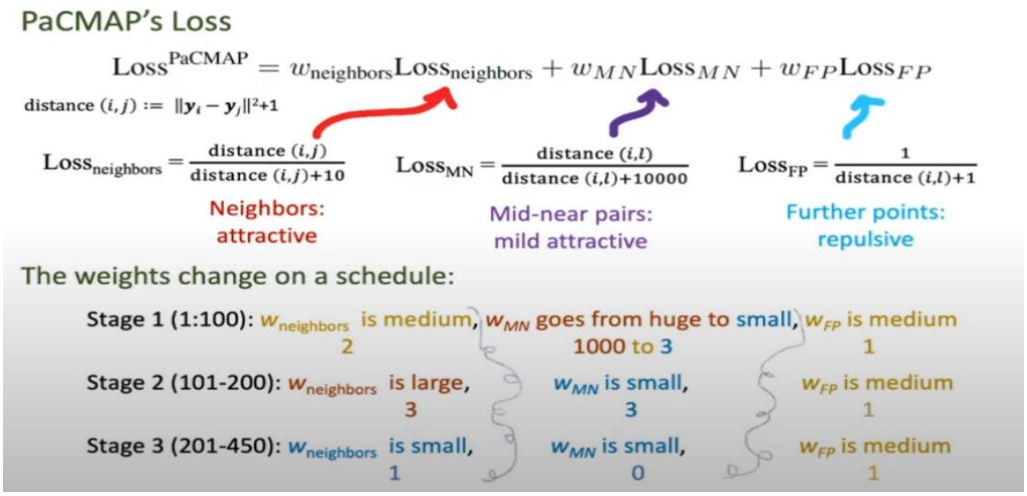
1. Near Pair(근접 쌍) : 두 데이터 포인트 간의 거리가 상대적으로 가까운 경우를 나타낸다. 이를 활용하여 데이터의 구조를 보존하면서 차원 축소를 진행한다.
2. Mid Near Pair (중간 근접 쌍) : Near Pair 보다는 조금 더 먼 거리에 있는 두 데이터 포인트를 나타낸다. 이는 데이터의 미세한 구조를 포착하고 보존하기 위해 활용된다.
3. Further Pair(먼 쌍) : 이는 상대적으로 먼 경우를 나타낸다.

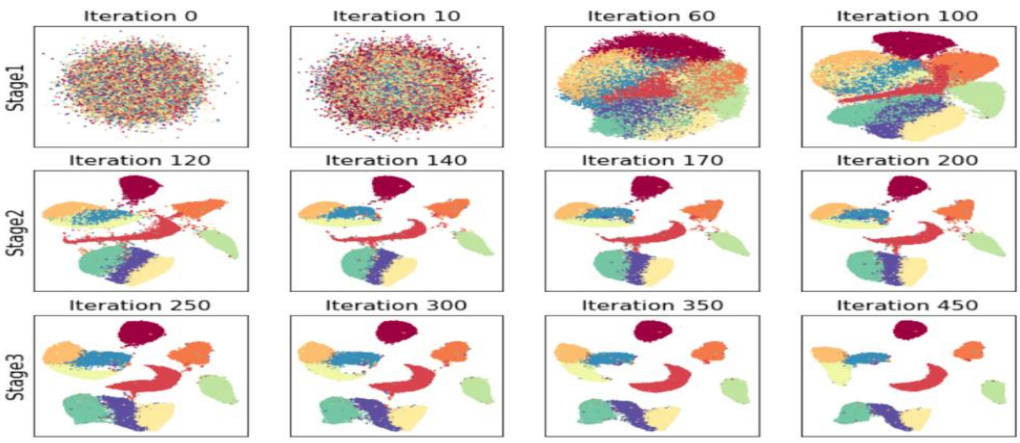
Loss를 최소화하기 위해, 위와 같이 3개의 stage로 나눠 학습을 진행하고, 각 stage 별로 가중치를 달리하여 local/global을 구분한다.





- UMAP 참고 사진(위) // PaCMAP 참고 사진(아래)





참고자료 : https://kwonkai.tistory.com/65 // https://slowsteadystat.tistory.com/30