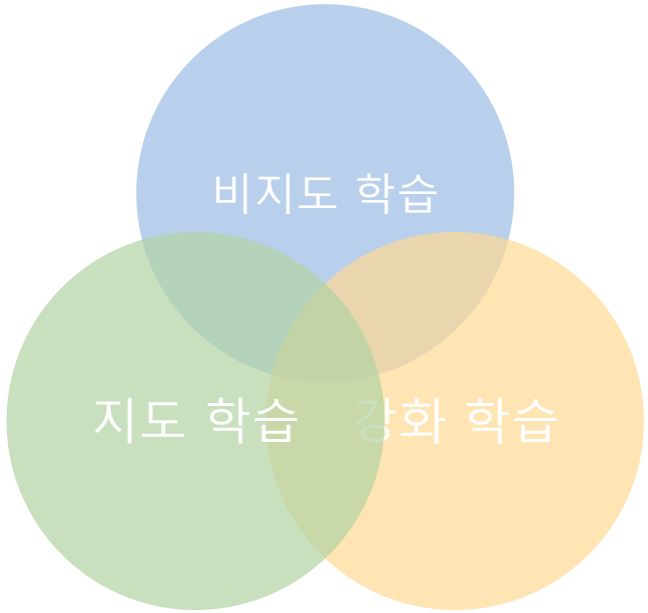
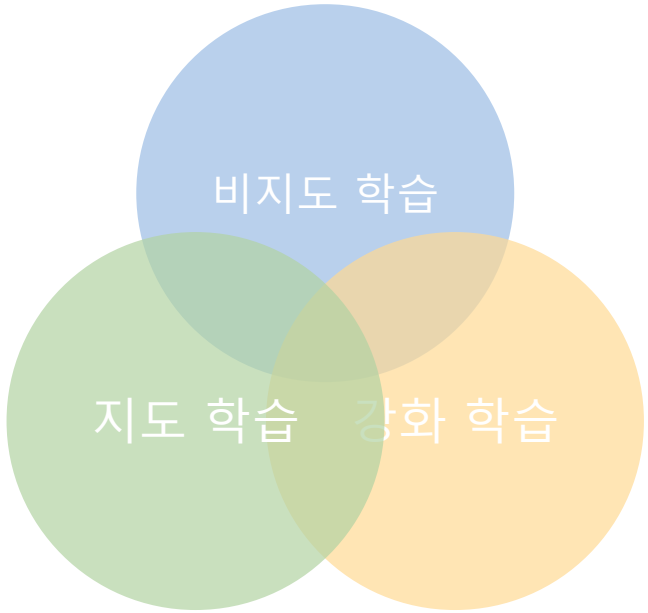


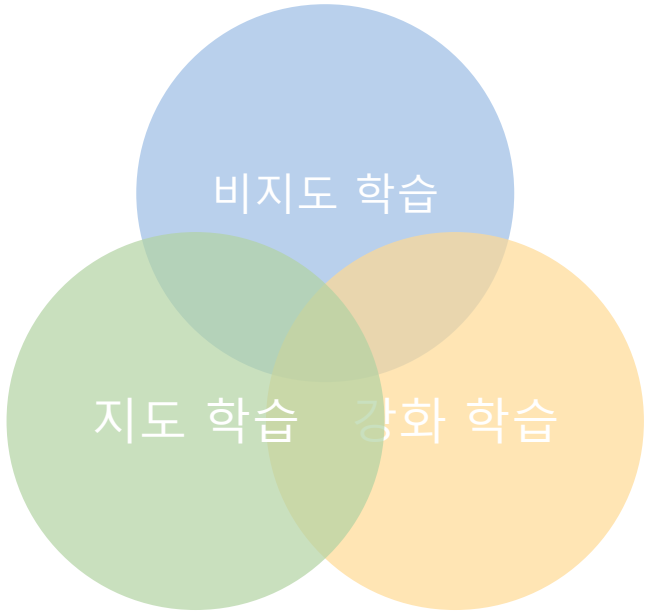
3가지 외 인공지능 학습: 자기지도학습

- 자기지도학습은 최소한의 데이터만으로 스스로 규칙을 찾아 분석하는 AI기술이다. 사람이 별도로 지도하지 않아도 기계가 스스로 대상을 인지하고 의미를 부여한다.
- 라벨링을 하나씩 붙이던 기존의 지도학습 방식에는 인공지능 기술이 발전하는데 한계가 있었다. 이는 딥러닝 커뮤니티에서 핵심 이슈였으며 이 문제점을 해결하기 위해 자기지도학습, 비지도 학습, 공백을 매우는 학습등(Learning to Fill In the Blank)이 필요하다.



3가지 외 인공지능 학습: 자기지도학습

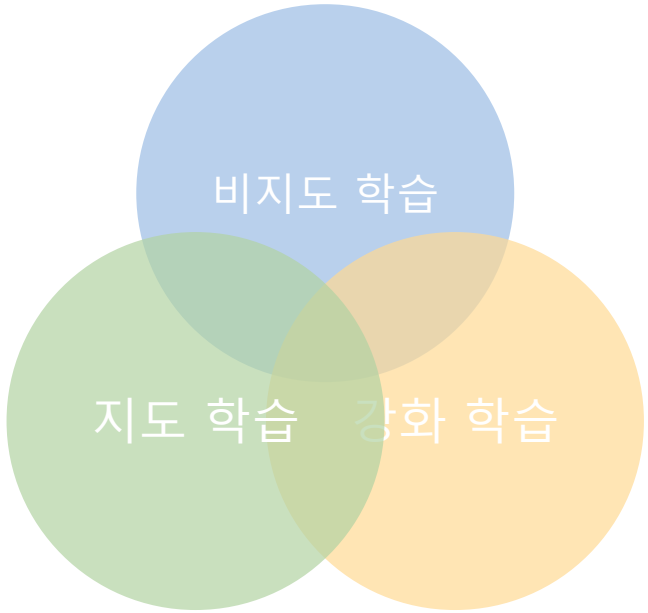
- 2014년 메리 미커의 인터넷 트렌드 리포트를 보면 매일 사람들이 트위터나 페이스북, 인스타그램에 업로드하는 이미지 개수는 18억개가 수준이다. 이를 1년 단위로 계산하면 6570억개인 반면 이미지넷 데티어세트는 1200만개 수준이다. 간단히 설명하면 연간 수천억개의 이미지 데이터가 쏟아지는데 레이블링된 데이터의 수는 1000만개 수준에 그치는 것이다. 이는 레이블링에 쓰이는 비용이 많이 들기 때문이다. 만약 레이블이 안 된 매년 쏟아지는 수천억개의 데이터를 활용하면 정말 놀라운 인공지능 발전, 딥러닝의 발전이 이뤄질 수 있을 것이다.



3가지 외 인공지능 학습: 준지도학습

(Semi – Supervised Learning)

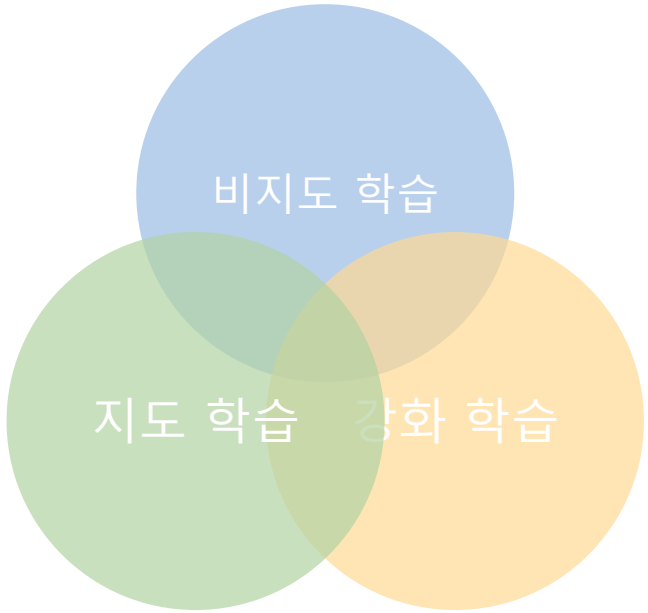
- 준지도학습 이란 기계학습 (머신 러닝) 의 한 범주로 목표 값이 표시된 데이터와 표시되지 않은 데이터를 모두 훈련에 사용하는 것을 말합니다.
- 지도학습은 정답(목표 값)이 태그 된 데이터를 학습하는 것으로 시작합니다.
- 이런 학습 과정을 거치면 가중치가 부여된 모델이 나오고 이 모델을 통해 태그가 되지 않은 비슷한 데이터가 입력 됐을 때 답변을 예측하게 됩니다.
- 준지도학습은 태그 된 데이터와 태그가 되지 않은 데이터 모두 활용합니다.



3가지 외 인공지능 학습: 준지도학습

(Semi – Supervised Learning)

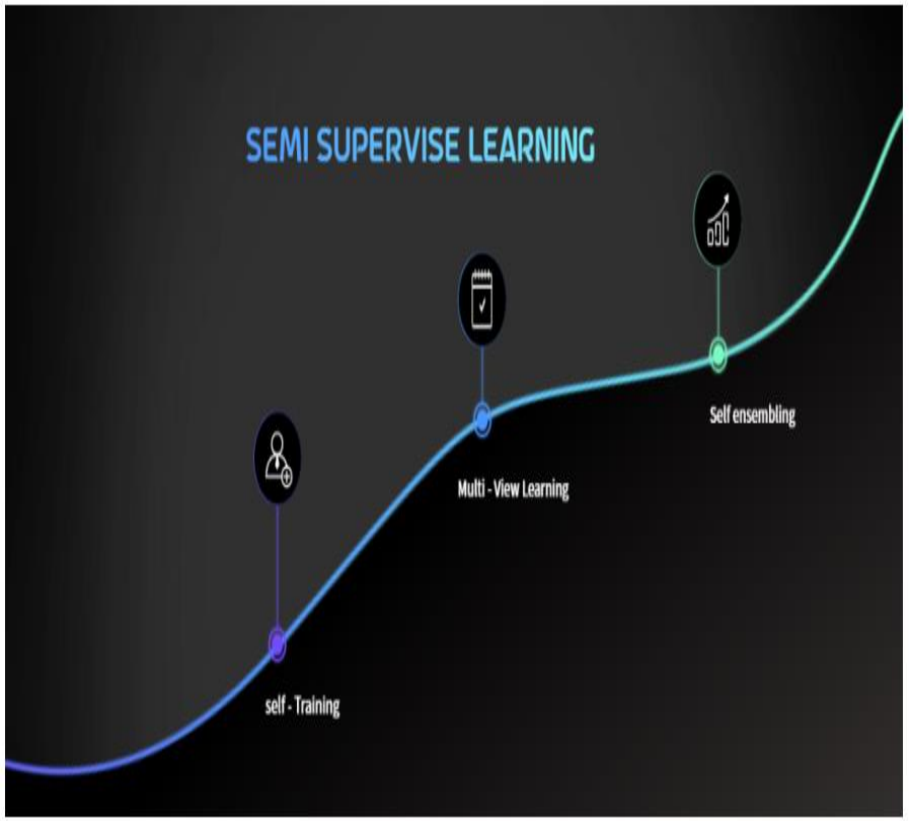
- 일반적인 경우 태그가 되지 않은 데이터는 오히려 모델의 정확성을 떨어트릴 수 있지만 알렉사는 태그가 되지 않은 데이터를 추가해 오히려 모델의 정확도를 개선 했습니다.
- 데이터에 태그를 하는 작업은 많은 시간과 비용이 들지만 항상 그런 것은 아닙니다.
- 즉, 많은 데이터를 갖고 있고 이 중 일부만 태그가 되어 있다면 준지도학습을 테스트 해 볼만 하다고 볼 수 있습니다.



3가지 외 인공지능 학습: 준지도학습 목표

(Semi - Supervised Learning)

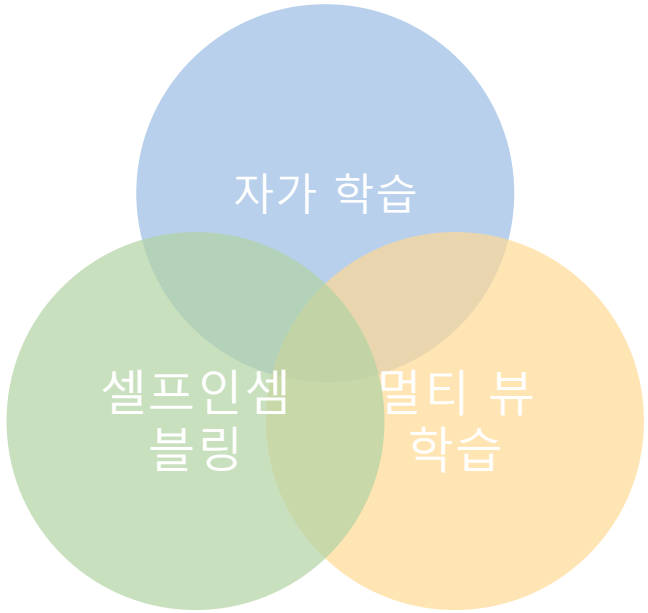
- 준지도학습의 목표는 간단합니다.
- 레이블이 달려있는 데이터와 레이블이 달려있지 않은 데이터를 동시에 사용해서 더 좋은 모델을 만드는 것을 의미합니다.



3가지 외 인공지능 학습: 준지도학습 언레이블 데이터

(Semi – Supervised Learning)

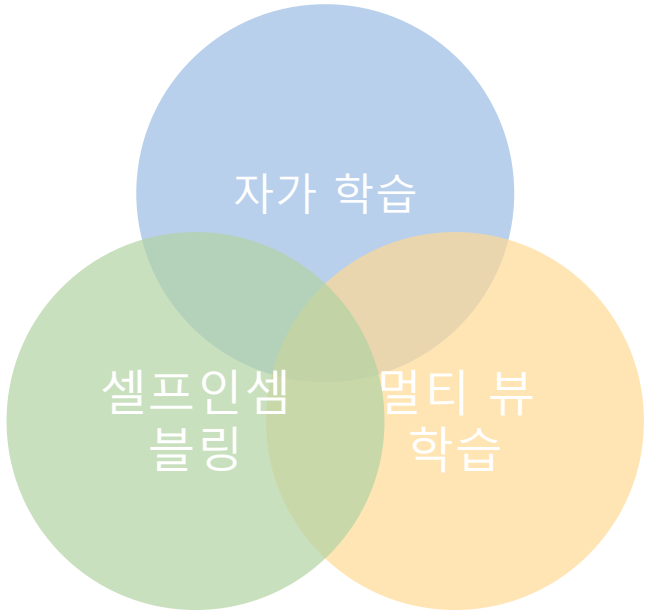
- 언제 언레이블 데이터들이 학습에 도움을 줄 수 있는 방법은 아래와 같습니다.
- 언레이블 데이터들의 분포가 만약 균등하다면 지도학습에 전혀 도움이 되지 않을 수 있습니다.
- 균등한 분포를 가지고 있는 언레이블 데이터를 아무리 더해줘도 기존 모델의 성능을 향상시킬 수 없습니다.
- 하지만 반대로 군집 형태라면 학습에 도움이 될 수 있습니다.
- 현실 세계의 데이터들은 클러스터 과정을 만족하는 경우가 대부분이기 때문에 최소한 준지도학습을 사용하게 되면 손해볼 일이 없다고 할 수 있습니다.



3가지 외 인공지능 학습: 준지도학습의 세가지, 자가학습

(Self - Training)

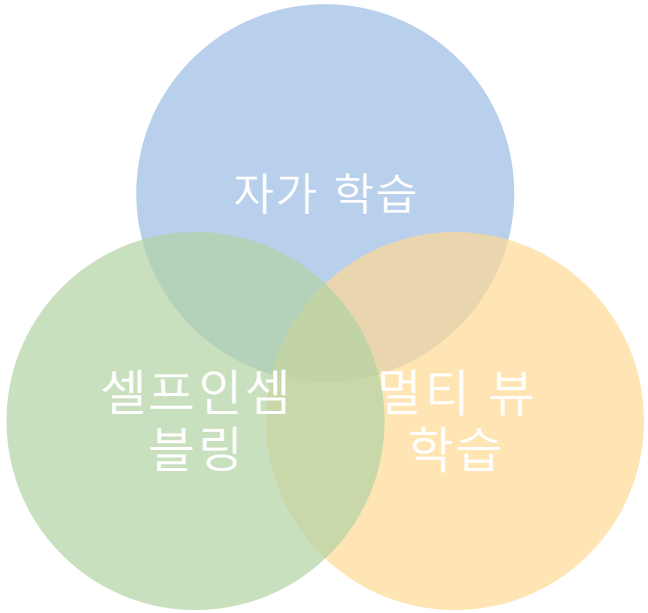
- 자가학습 라벨된 데이터 셋을 추가하는데 라벨되지 않은 데이터에 대한 모델의 자제 예측을 이용합니다.
- 기본적으로 사용자가 예측의 신뢰성을 결정하는 기준점을 설정하는데 보통 0.5 혹은 그 이상입니다.
- 이보다 높으면 에측을 신뢰하고 이를 라벨된 데이터 세셍에 추가한다는 의미 입니다.
- 이러한 학습 작업은 신뢰할 수 있는 예측이 없을 때 까지 계속해서 반복합니다.
- 자가 학습은 반쯤 성공했다는 평가를 받고 있음을 의미합니다.



3가지 외 인공지능 학습: 준지도학습의 세가지, 멀티 뷰 학습

(Multi - View Learning)

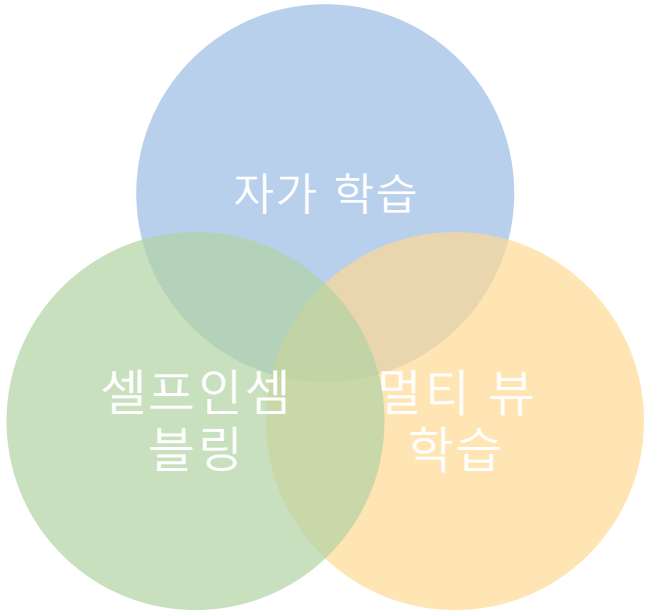
- 멀티 뷰 학습은 데이터의 여러 측면에서 여러 모델을 학습합니다.
- 여기에는 다른 기능 셋, 다른 모델 아키텍처 또는 다른 데이터 서브 셋이 포함이 될 수 있습니다.
- 멀티 뷰 학습은 알고리즘이 매우 다양하지만 가장 널리 알려진 것이 트라이 학습(Try - Training) 입니다.
- 트라이 학습은 기본적으로 3개의 모델을 만들 수 있습니다.
- 2개의 모델이 데이터 라벨에 합의할 때마다 이 라벨이 세 번째 모델에 추가됩니다.
- 자가학습처럼 더는 새로 추가할 라벨이 없을 때는 작동을 멈추게 됩니다.



3가지 외 인공지능 학습: 준지도학습의 세가지, 셀프인셈블링

(Self - ensembling)

- 셀프 인셈블링은 일반적으로 다양한 설정의 단일 모델을 사용합니다.
- 사다리형 회로망 모델에서 잘 정제된 사례에 대한 예측은 임의의 불안정한 사례에 대한 프록시 라벨로 사용됩니다.
- 이 방식을 기준으로 노이즈를 견딜 수 있는 기능을 개발할 수 있습니다.



3가지 외 인공지능 학습: 준지도학습은 유사도를 이용

- 준지도학습은 임베딩 값의 유사도를 이용합니다.
- 이와 달리 준지도학습 모델의 기본 아이디어는 같은 class에 속하는 이미지를 구별하는 근거인 임베딩 값이 유사할 것이라는 가정입니다.
- 임베딩 값의 유사도를 이용해 더 정확한 임베딩 값을 만드는 모델을 제안합니다.
- 학습 데이터 관련해서 일반적인 딥러닝 모델이 정답이 달린 레이블 데이터만 이용하는 것과 달리 준지도학습 모델은 정답이 없는 지정되지 않은 언레이블 데이터 모두를 이용해야 합니다.
- 예를 들어, 필기체 숫자 이미지를 예측하는 모델 학습 시 레이블 데이터 3장과 데이터 3장을 입력해 사용한다면 딥러닝 모델은 각각의 이미지에대한 일정한 길이의 숫자로 구성된 임베딩 값을 생성하게 됩니다.

딥러닝은 Deep Learning
임베딩을 요구하다.

3가지 외 인공지능 학습: 일반적인 딥러닝은 정확한 임베딩 값 요구

- 일반적인 딥러닝 모델을 학습하기 위해서는 분류해야 할 각 대상에 대해 수집, 수백만 장의 이미지를 정답과 함께 입력해야 됩니다.
- 데이터가 입력되면 딥러닝 모델은 입력된 이미지 데이터의 RGB 값을 이용하여 각 대상을 구별할 수 있는 숫자 집합을 생성하는 데 일반적으로 적게는 수백 개에서 많게는 수천 개의 숫자 집합을 생성합니다.
- 이때 생성된 숫자 집합을 벡터 혹은 임베딩 값이라 부릅니다.
- 딥러닝 모델은 생성된 임베딩 값을 이용하여 이미지가 무엇인지를 예측하고 예측된 결과와 정답을 비교한 후 더 정확한 예측을 할 수 있도록 모델을 개선합니다.
- 즉, 각각의 이미지를 구별하는 근거인 임베딩 값이 정확해지도록 모델이 학습하게 되어있습니다.