Deep Learning with Big Data : An Emerging Trend

2019 19th International Conference on Computational Science and Its Applications (ICCSA)

성균관대학교
데이터사이언스융합학과 석사과정
2019712688 조수현

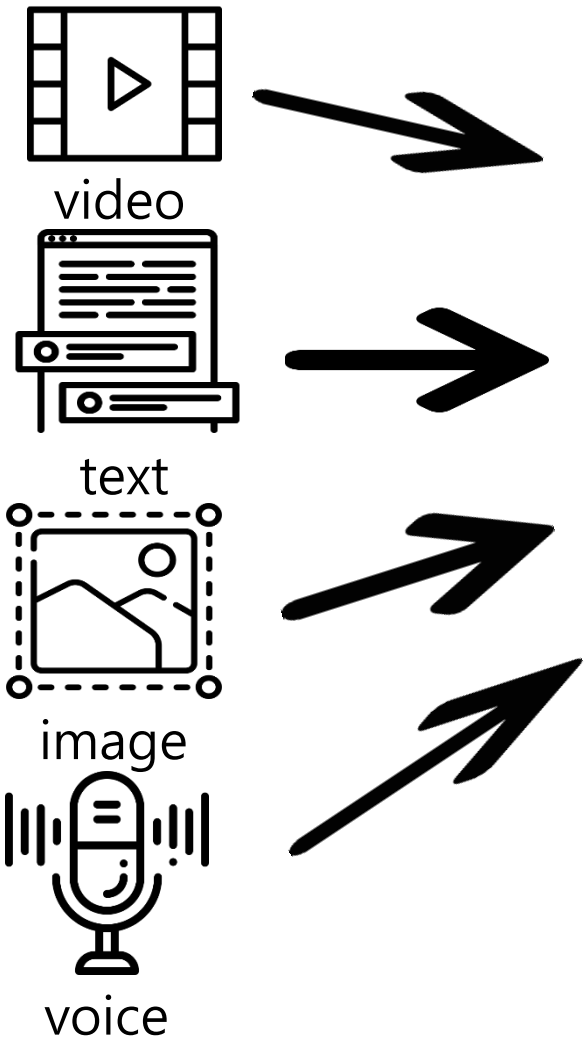
CONTENTS

1. Introduction

1. 기존 데이터 분석 방식

2. 분석 방식의 변화

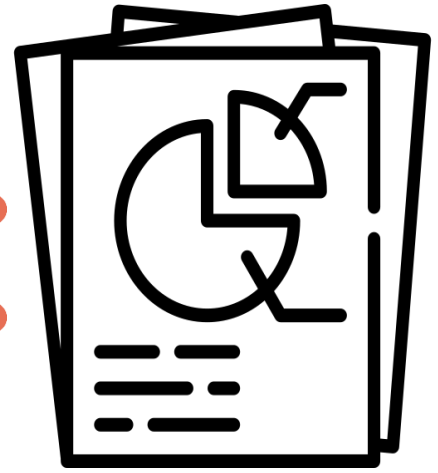
. 기존 데이터 분석 방식



Data의 빠른 증가

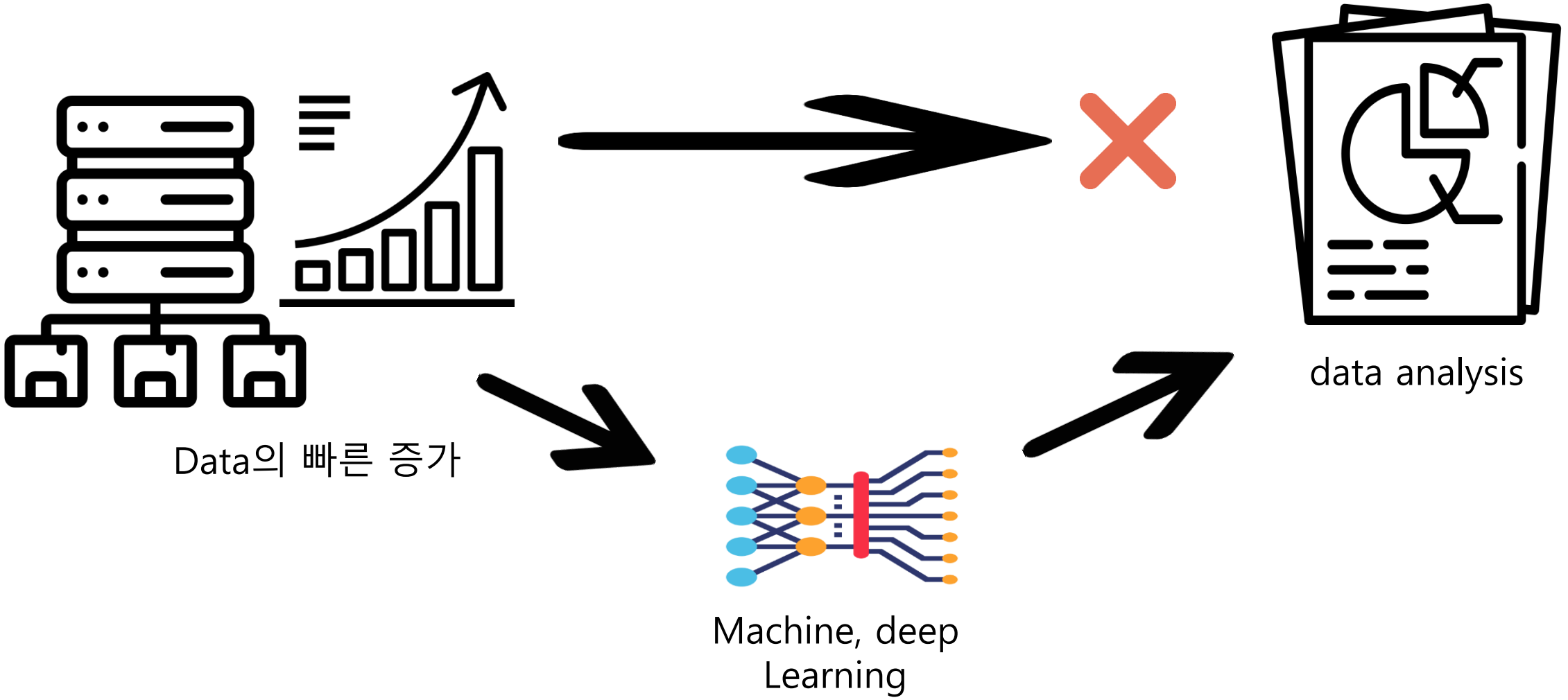
기존 방식으로는 데이터를
분석할 수 없음

1. 다양한 종류의 데이터
2. 데이터의 노이즈
3. 데이터 크기의 증가



data analysis

. 분석 방식의 변화



2. Overview Of Deep Learning Architecture

1. Auto-Encoders (AEs)
2. Restricted Boltzmann Machine (RBM)
3. Deep Belief Network (DBN)
4. Generative Adversarial Network (GAN)
5. Convolutional Neural Networks (CNN)
6. Deep Boltzmann Machine (DBM)
7. Deep Stacking Network (DSN)
8. Recurrent Neural Network (RNN)

. Auto-Encoders (AEs)

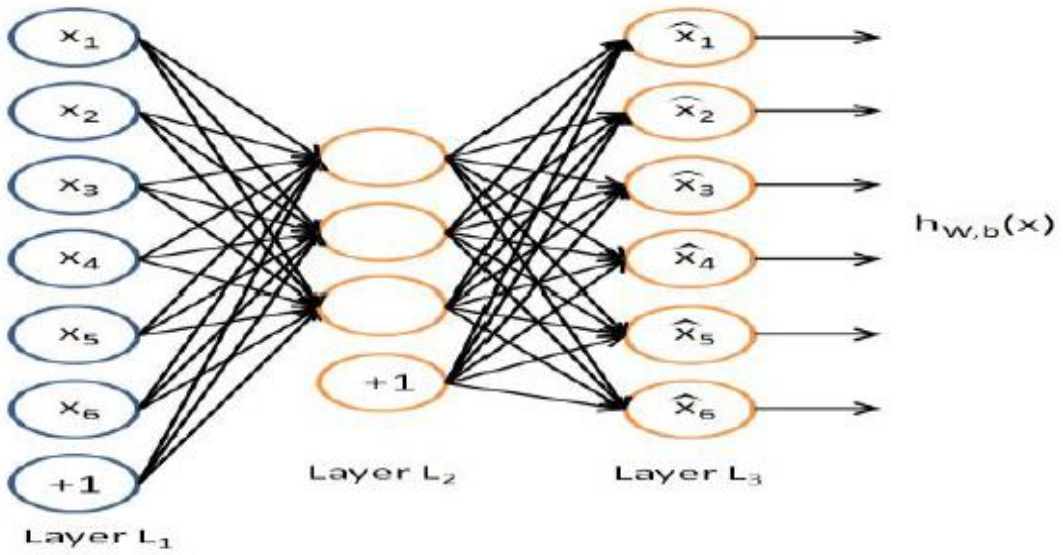


Fig. 1. The Auto-Encoder neural network in which the output is similar to the input [24]

- 설명

- 출력 계층에서 입력을 재구성하기 위해 생성하는 방법
- 이미지 색채, 치수 감소, 특징 변형, 워터마크 제거 및 변성 영상에 사용

- 특징

- 비지도 학습
- 단층 학습 알고리즘

. Restricted Boltzmann Machine (RBM)

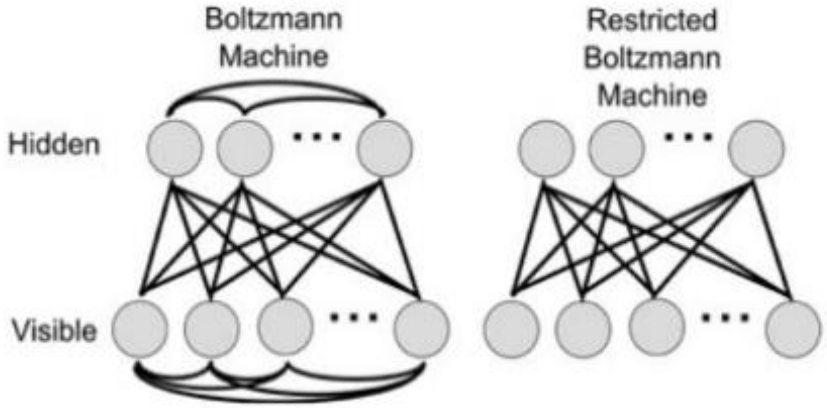
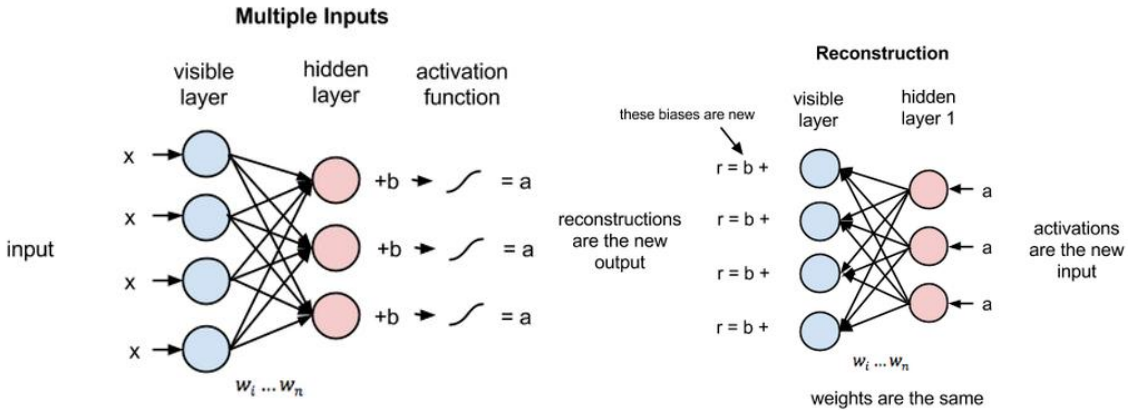


Fig. 2. The Restricted Boltzmann Machine [17]



- 설명

- 차원 감소, 분류, 선형 회귀 분석, 협업 필터링(collaborative filtering), 특징값 학습(feature learning) 및 주제 모델링(topic modelling)에 사용할 수 있는 알고리즘으로 Geoff Hinton이 제안한 모델
- RBM의 구조는 상대적으로 단순한 편
- RBM은 자체적으로도 사용할 수 있지만 다른 심층 신경망의 학습을 돕기 위해 쓰이기도 한다.
- RBM은 *심층 신뢰 신경망*(DBN:Deep Belief Network)을 구성하는 요소로 쓰인다.

- 특징

- 두 개의 층(입력층 1개, 은닉층 1개)으로 구성
- 단층 학습 알고리즘
- 입력층의 노드는 데이터를 입력으며 입력받은 데이터를 은닉층에 얼마나 전달할 것인지를 확률에 따라 결정 한다.
- 비지도 학습

https://steemit.com/kr/@yoonheeseung/2-restricted-boltzmann-machines-rbms

. Deep Belief Network (DBN)

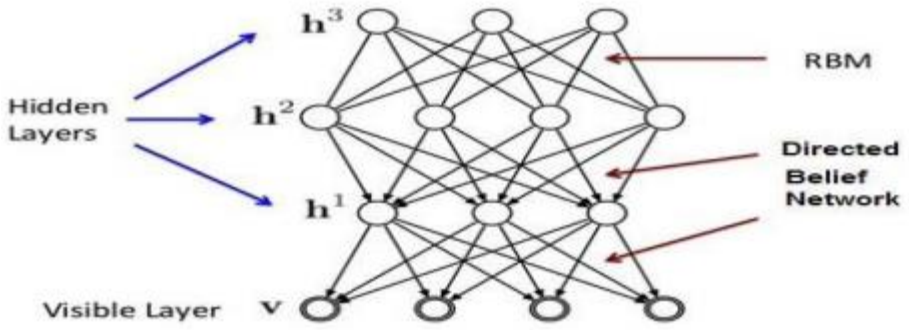
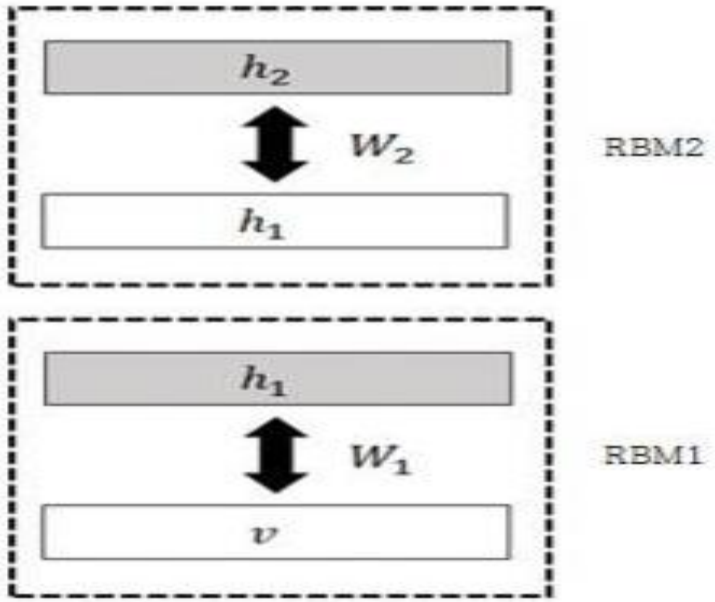


Fig. 3. The Deep Belief Network [29]



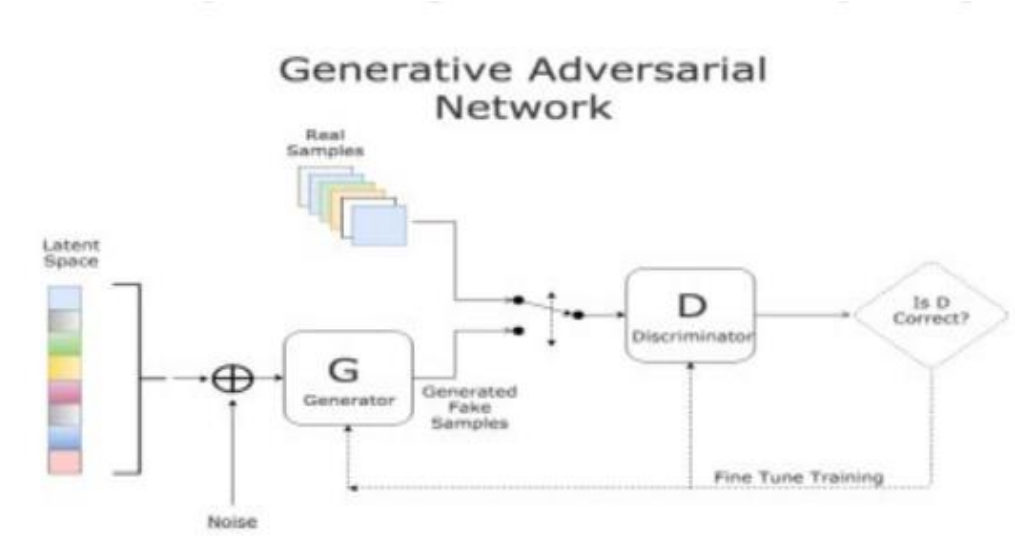
- 설명

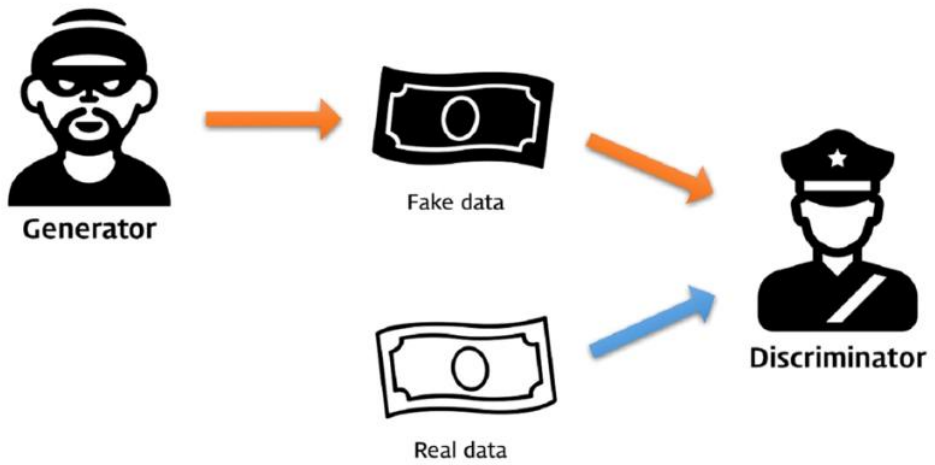
- RBM을 이용해서 MLP(Multilayer Perceptron)의 Weight를 input 데이터들만 보고 Pretraining 시켜서 학습이 잘 일어날 수 있는 초기 세팅을 하는 방법
- Pretraining을 통해 초기 가중치를 학습한 후 , backpropagation이나 다른 판별 모델을 위한 알고리즘을 통해 가중치를 미세 조정할 수 있다.

- 특징

- 학습데이터가 적을 때 유용
- 단층 학습 알고리즘
- 사전 훈련된 RBM을 층층이 쌓아 올려 만듦
- 비지도 학습

. Generative Adversarial Network(GAN)





- 설명

- 제로섬 게임 틀 안에서 서로 경쟁하는 두 개의 신경 네트워크 시스템에 의해 구현
- 2014년에 이안 굿펠로우에 의해 발표되었다.
- 원 데이터가 가지고 있는 확률분포를 추정하도록 하고, 인공 신경망이 그 분포를 만들어 낼 수 있도록 한다

- 특징

- 비지도 학습일반적인 머신 러닝, 혹은 딥 러닝 모델과는 달리 명확한 평가의 기준이 없다

. Convolutional Neural Networks (CNN)

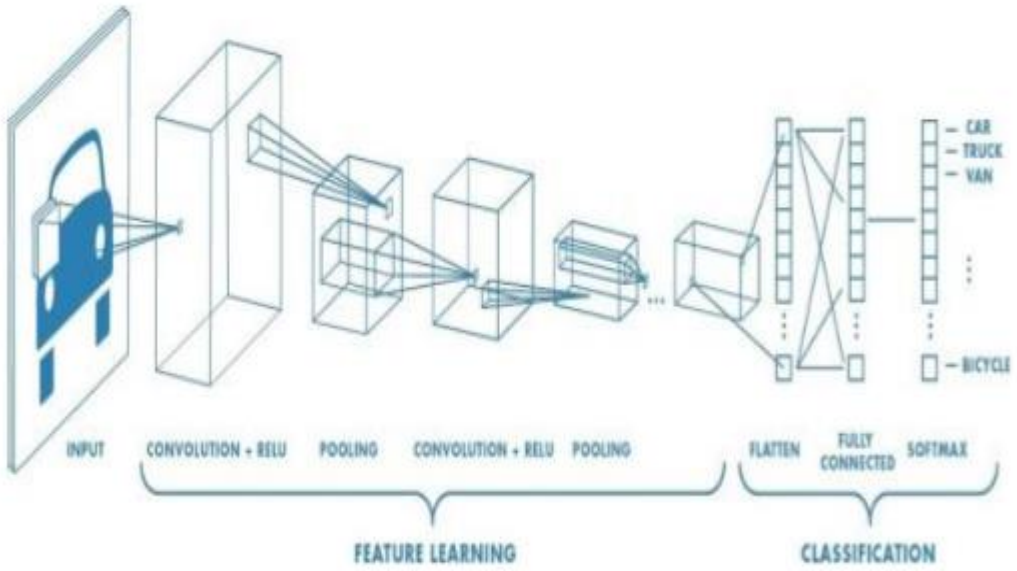


Fig. 5. Convolutional Neural Networks [52]

- 설명

- 공간 정보를 유지하면서 인접 데이터와의 특징을 효과적으로 인식
- 복수의 필터로 데이터의 특징 추출 및 학습
- 추출한 데이터의 특징을 모으고 강화하는 Pooling 레이어
- 필터를 공유 파라미터로 사용하기 때문에, 일반 신공망에 비해 학습 파라미터가 매우 적음

- 특징

- **Locality :** CNN에서 집중하는 부분은 object의 local structure
- Translation Invariance : Object에 대한 인식은 location과 independent
- Weight sharing : feature map을 만드는데는 동일한 filter를 사용하기 때문에, weight를 학습하는 filter의 경우 다들이 weight를 공유

. Deep Boltzmann Machine (DBM)

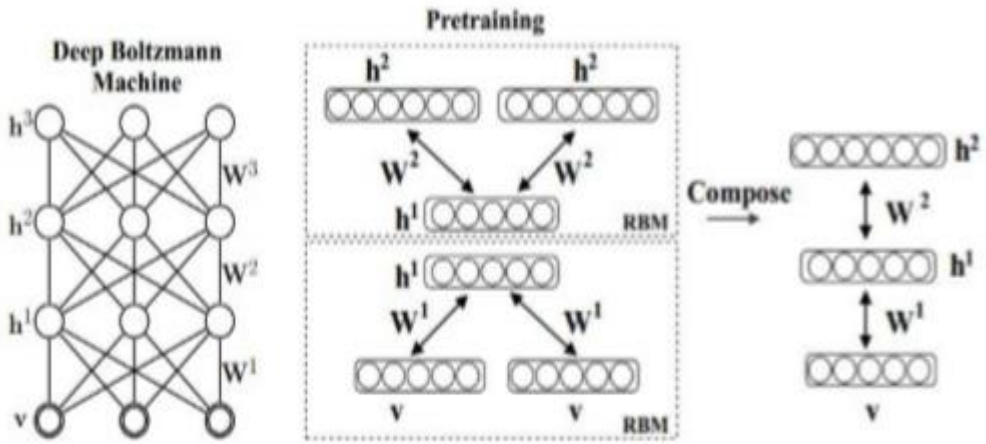


Fig .6. The Deep Boltzmann Machine and its Pre-training procedure [30]

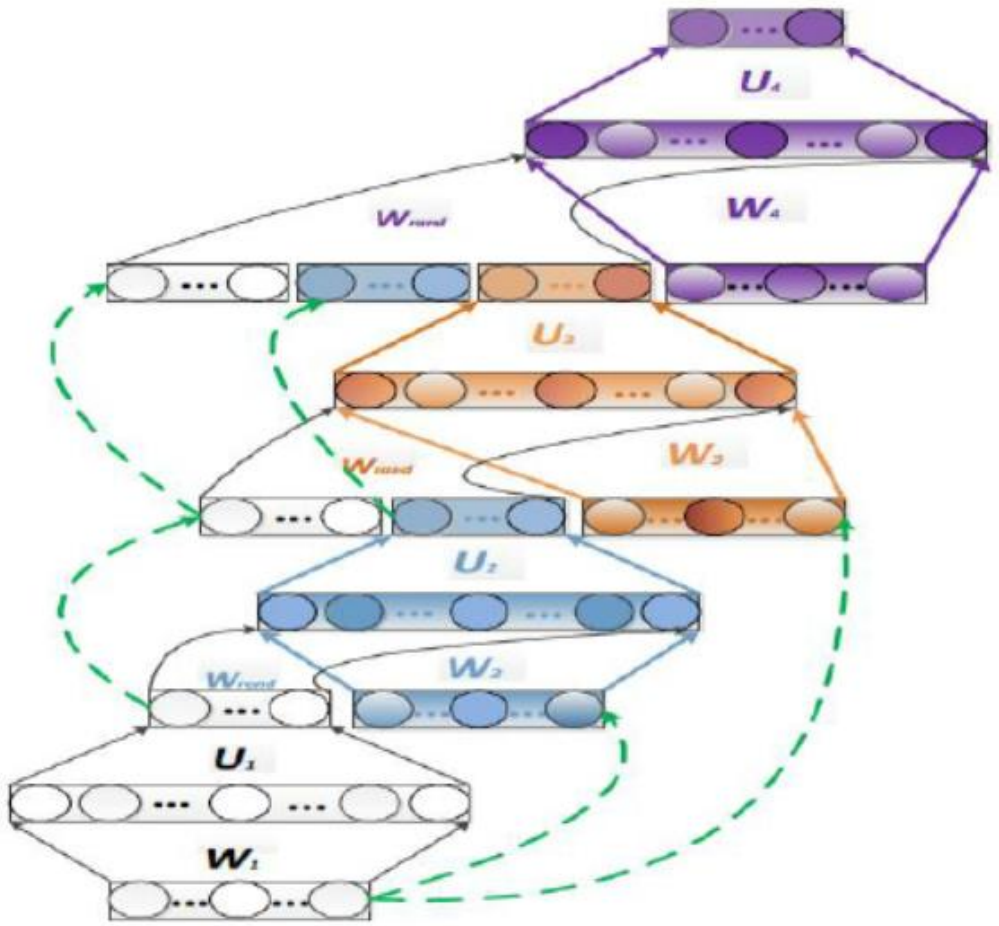
- 설명

- DBN과 마찬가지로 양방향으로 추론과 학습을 함으로써 모호하고 복잡한 입력 구조의 표현을 더 잘 학습할 수 있다.
- 레이블이 없는 다량의 입력 데이터를 사용하여 학습된 표현을 레이블된 소량의 데이터를 사용하여 미세 조정할 수 있다.
- RBM에서는 매우 간단한 추론 규칙이 있지만. DBM의 추론은 근사치
- 완전 무 방향 연결

- 특징

- 비지도 학습
- RBM과 마찬가지로 DBM에는 계층 내 연결이 없다. 연결은 인접 레이어의 단위 사이에만 존재
- 반지도 학습에 유용

. Deep Stacking Network (DSN)



A Deep Stacking Network Architecture

- 설명

- 모듈은 순서대로 학습되어 더 낮은 층의 가중치가 각 단계에서 계산된다.
- 각 블럭은 같은 최종 레이블 클래스 y의 평가치를 계산하고, 각 평가치는 본래의 입력 x와 연결되어 다음 블럭에 대한 확장된 입력값을 구성한다.
- 각 블럭은 단일 은닉 계층을 갖는 단순화된 다계층 퍼셉트론(multi-layer perceptron, MLP)으로 이루어져 있다.

- 특징

- CPU 클러스터 혹은 GPU 클러스터 상에서 배치 모드로 병렬 처리된다.
- 순수한 discriminative task에서 DSN은 일반적인 DBN보다 더 성능이 좋다.

. Recurrent Neural Network (RNN)

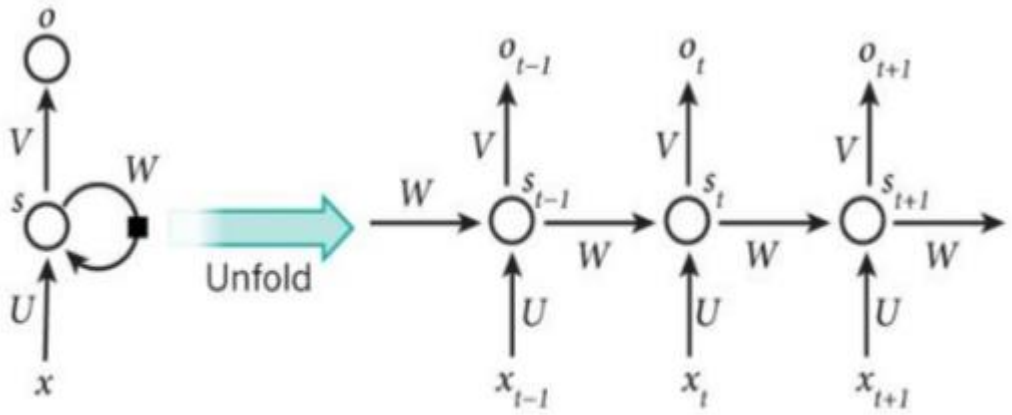
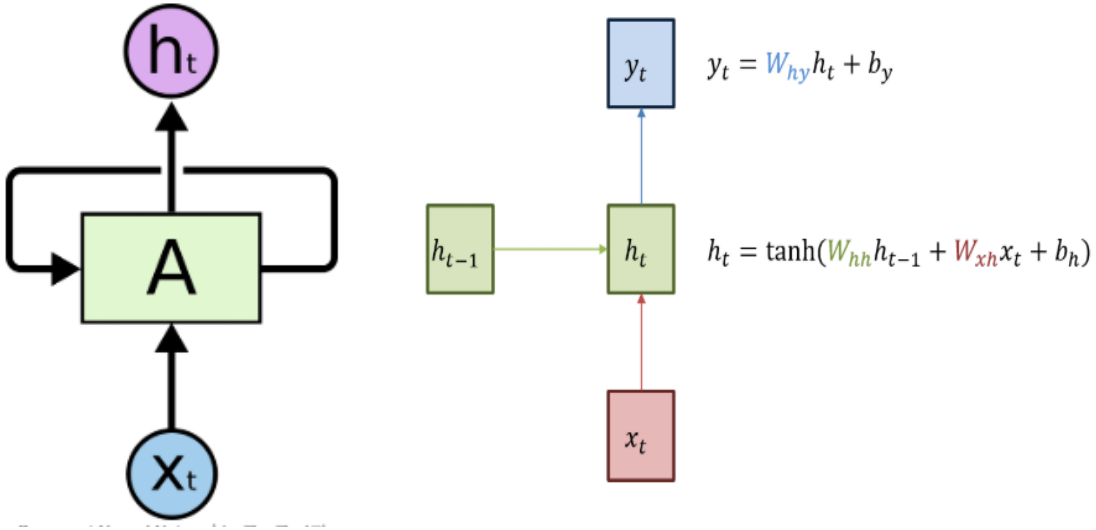


Fig. 7. Recurrent Neural Networks [53]



- 설명

- 시퀀스 데이터를 모델링 하기 위해 등장
- 히든 노드가 방향을 가진 엣지로 연결돼 순환구조를 이루는 (directed cycle) 인공신경망의 한 종류

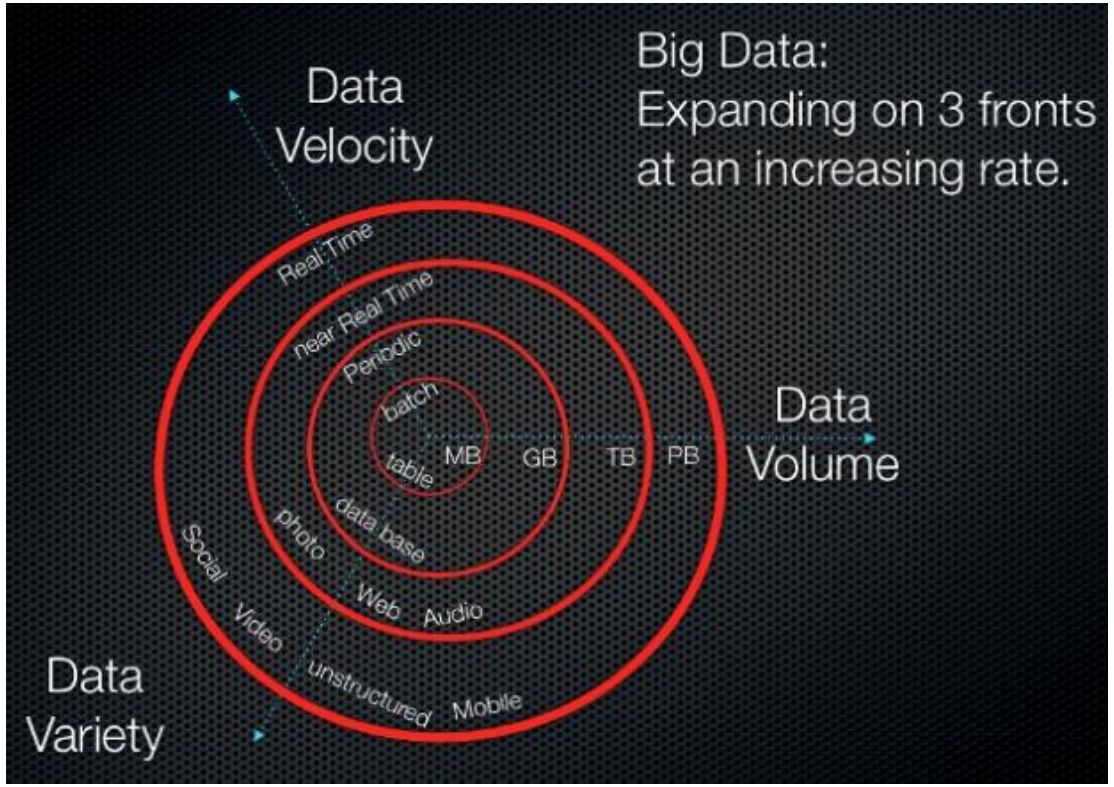
- 특징

- 시퀀스 길이에 관계없이 인풋과 아웃풋을 받아들일 수 있는 네트워크 구조
- One or Many to One or Many 총 5가지(Many to Many는 두가지)의 구조가 있음

3. DEEPLEARNING FOR BIGDATA ANALYTICS

1. 3 V’s of Big Data
2. 4 V’s of Big Data
3. 5 V’s of Big Data
4. 10 V’s of Big Data
5. 42 V's of Big Data and Data Science

. 3 V′s of Big Data



- Velocity

정보의 양이 많아지는 만큼 데이터의 신뢰성이 떨어지기 쉽다. 이러한 측면에서 새로운 속성인 정확성(Veracity)이 제시되고 있다.

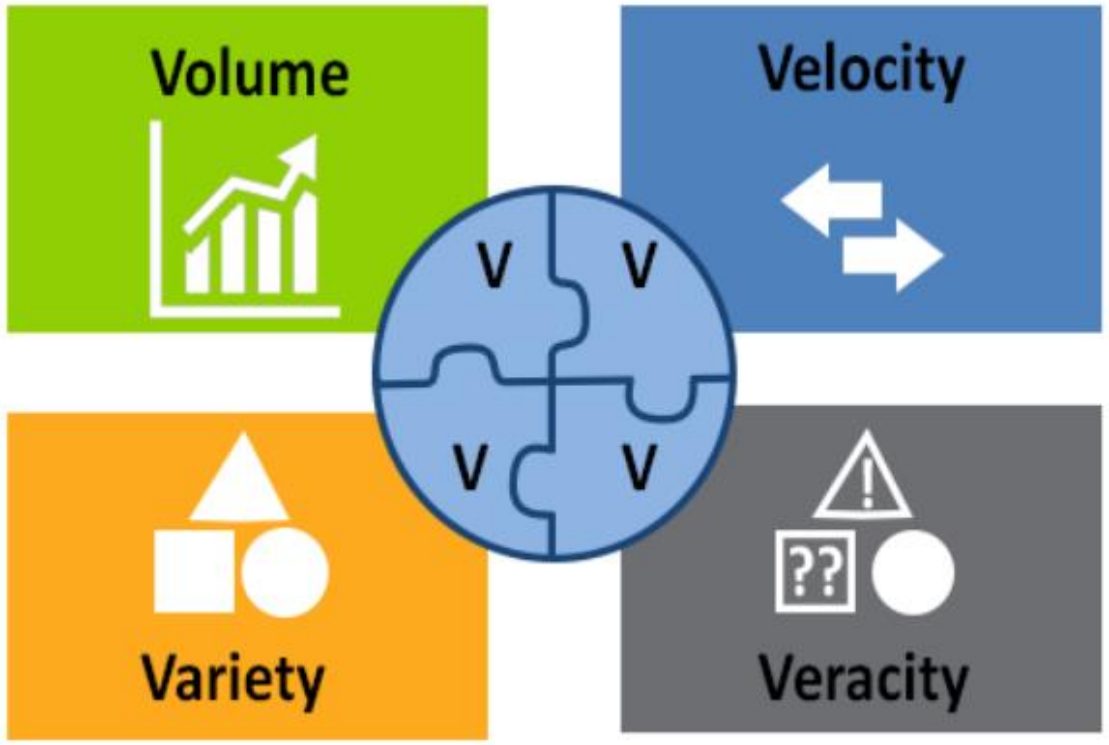
- Volume

대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성, 데이터는 매우 빠른 속도로 생산된다.

- Variety

정형, 비정형, 반정형 등의 다양한 종류의 데이터를 의미

. 4 V′s of Big Data



- Velocity

정보의 양이 많아지는 만큼 데이터의 신뢰성이 떨어지기 쉽다. 이러한 측면에서 새로운 속성인 정확성(Veracity)이 제시되고 있다.

- Volume

대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성, 데이터는 매우 빠른 속도로 생산된다.

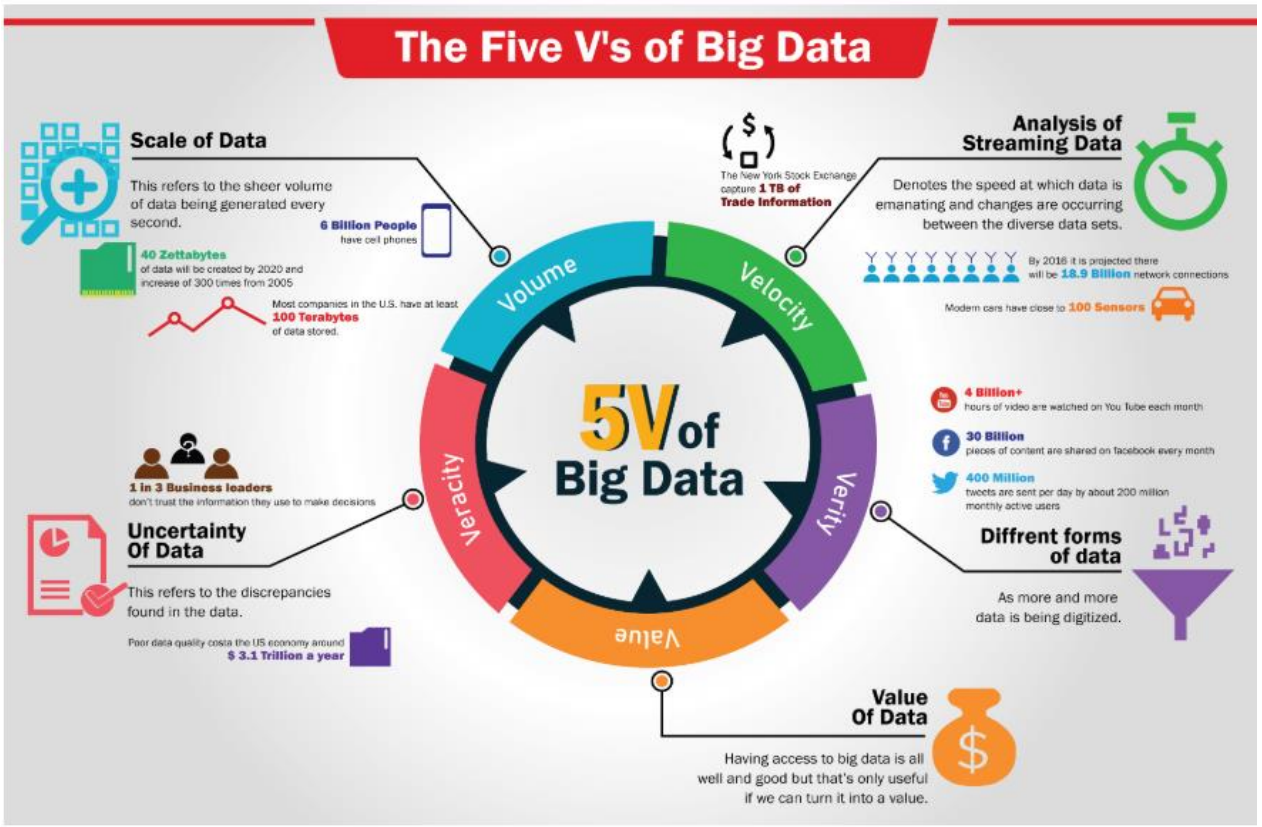
- Variety

정형, 비정형, 반 정형 등의 다양한 종류의 데이터를 의미

- Veracity

가짜 정보들 사이에서 정확한 정보를 사용하는 것
분석방법, 데이터의 오류제거 통한 신뢰할 수 있는 결과 도출이 중요

. 5 V′s of Big Data



- Velocity
- Volume
- Variety
- Veracity

- Value

다양하고 대량의 데이터간의 상호 연관성 분석과 의미 추출을 통하여 조직의 의사결정에 가치를 더해주는 결과 도출 지향

ZETAWIKI 참조

. 10 V′s of Big Data

Fig. 8. 10V's 0f Big Data [55]

Velocity, Volume, Variety, Veracity, Value

Validity : 데이터가 의도 한 용도에 대해 얼마나 정확하고 정확한지 나타낸다

Vulnerability : 빅 데이터는 새로운 보안 문제를 야기한다.

Volatility : 데이터가 더 이상 관련이 없거나 역사적이거나 유용하지 않은 것 -> 얼마나 오래 보관할지

Visualization : 빅 데이터의 시각화

Variability : 데이터의 불일치, 상치 탐지 방법으로 이를 찾아야한다.

. 42 V′s of Big Data

Vagueness	Validity	Valor	Value	Vane	Vanilla	Vantage
Variability	Variety	Varifocal	Varmint	Varnish	Vastness	Vaticination
Vault	Veer	Velocity	Veil	Venue	Veracity	Verdict
Versed	Version Control	Vet	Vexed	Viability	Vibrant	Victual
Viral	Virtuosity	Viscosity	Visibility	Visualization	Vivify	Vocabulary
Vogue	Volatility	Volume	Voodoo	Voyage	Vulpine	Voice

설명 : https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

4. APPLICATIONS OF DEEPLEARNING IN BIGDATA ANALYTICS

1. Semantic Indexing
2. Conducting Discriminative Tasks
3. Semantic Image and Video Tagging
4. Social Targeting

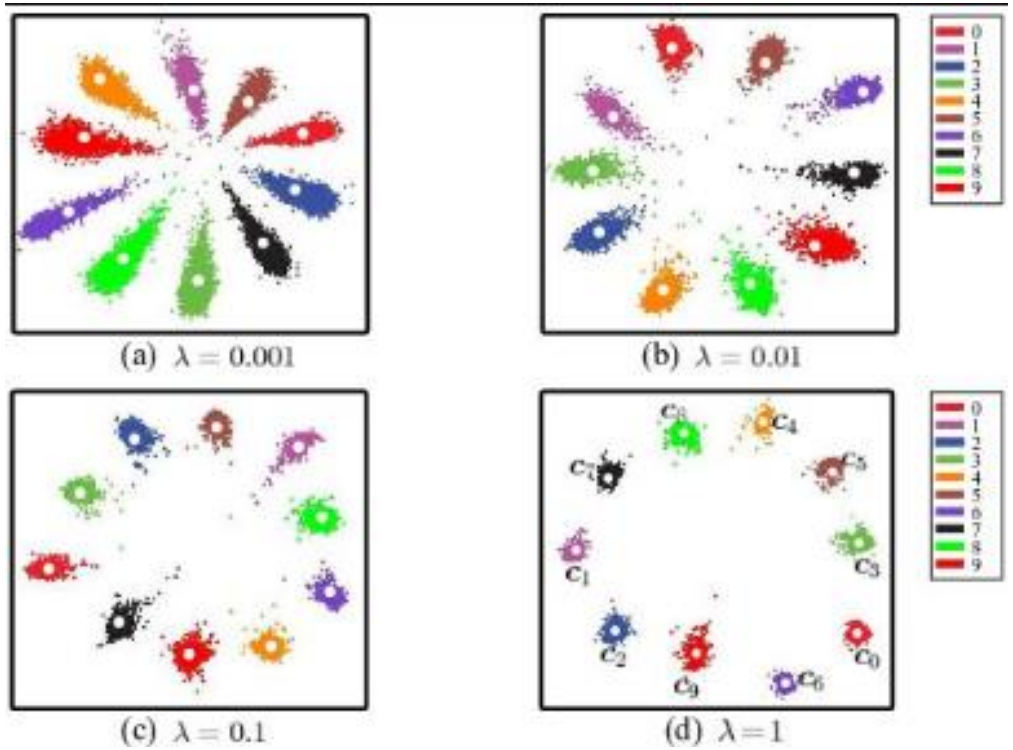
. Semantic Indexing

정보 검색은 Big Data 분석에서 유명한 필드 중 하나.

자연어 뿐만 아니라.

비디오, 이미지, 오디오, 사진 등에서도 검색

. Conducting Discriminative Tasks



선형 모델을 이용한 접근 방식을 이용하면 두 가지 장점이 있다.

1. Deep Learning을 사용하여 특징을 선택하여 데이터 분석에 비선형을 추가함 으로써 차별적 연산을 사용

2. 선택된 속성에 선형 분석 모델을 적용하는 것이 계산에 더욱 효율적

(좌 사진 판별 작업을 위한 선형 분석)

. Semantic Image and Video Tagging

- 이미지 태깅

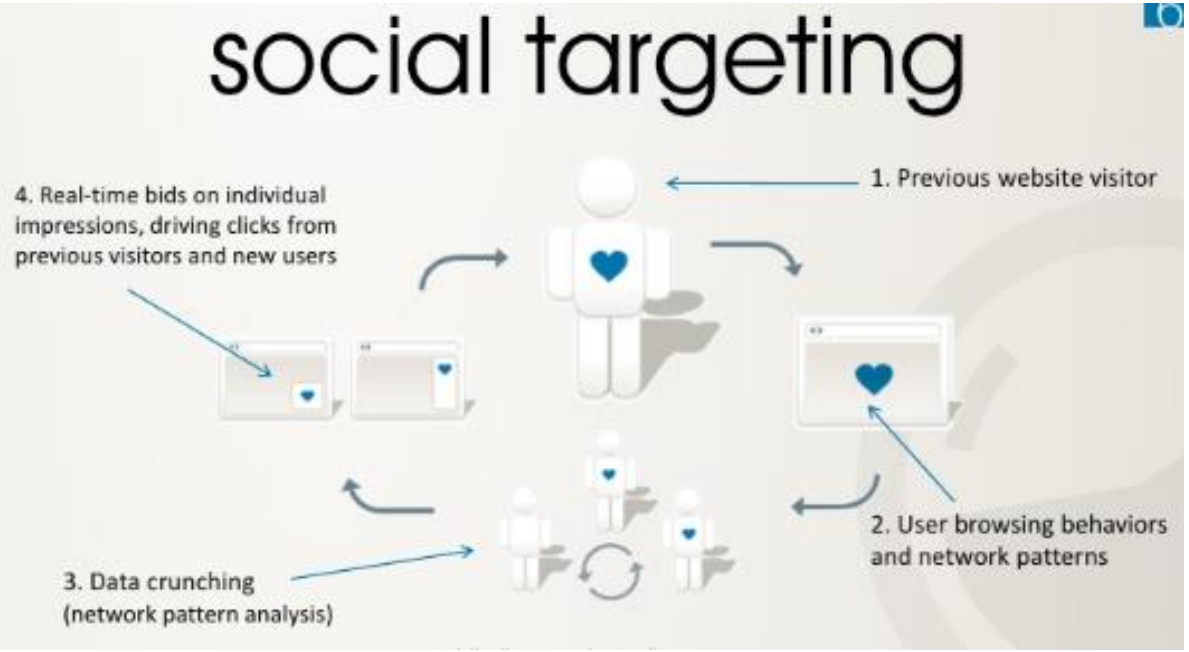
Deep Learning은 유형의 비디오 데이터 태깅 및 액션 장면을 인식.

- 시맨틱 태깅

텍스트를 분석하고 개념을 추출하며 주제, 키워드 및 중요한 관계를 식별하고 유사한 개체명을 찾음

BD Analytics에서 좋은 데이터 결과를 배우고 다른 복잡한 문제를 달성

. Social Targeting



- 소셜 타겟팅

소셜미디어 이용자들이 창출한 대화, 포스트, 코멘트 등을 분석하여 이용자들을 적절히 세분화(segment)하는 것

소셜미디어가 대중의 목소리를 듣고 기업과 소비자 간의 상호작용을 제공하는 중요한 수단으로 부상하고 있음

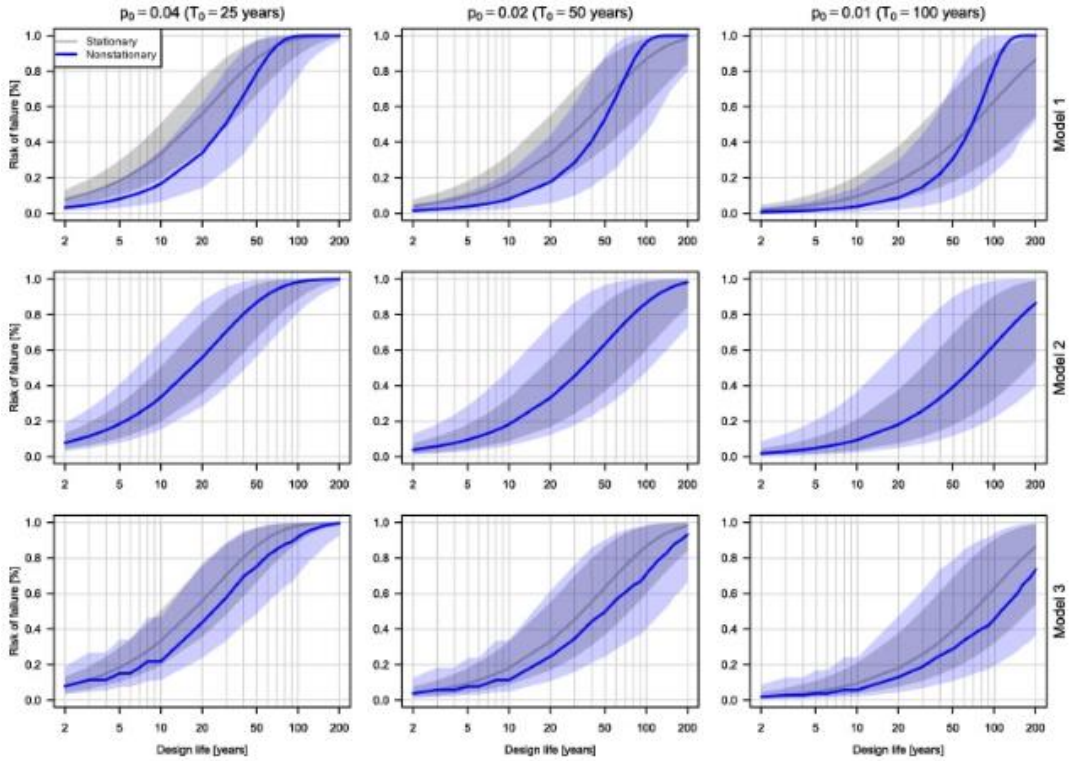
각 기업들에게 소셜데이터를 정확히 해석하는 것은 현재의 고객을 이해하는 것뿐 아니라 새로운 고객을 획득하기 위한 의미

https://www.slideshare.net/mac_wyszynski/gifversion-101122094132phpapp01

5. DEEPLEARNING APPLICATIONS FOR BIGDATA ANALYTICS

1. Real-time Non-stationary Data
2. High-dimensional Data
3. Data parallelism
4. Multimodal Data
5. Large-scale models

. Real-time Non-stationary Data



- 설명

실시간 데이터는 고속 데이터 생성을 나타내며 즉시 처리해야 한다.

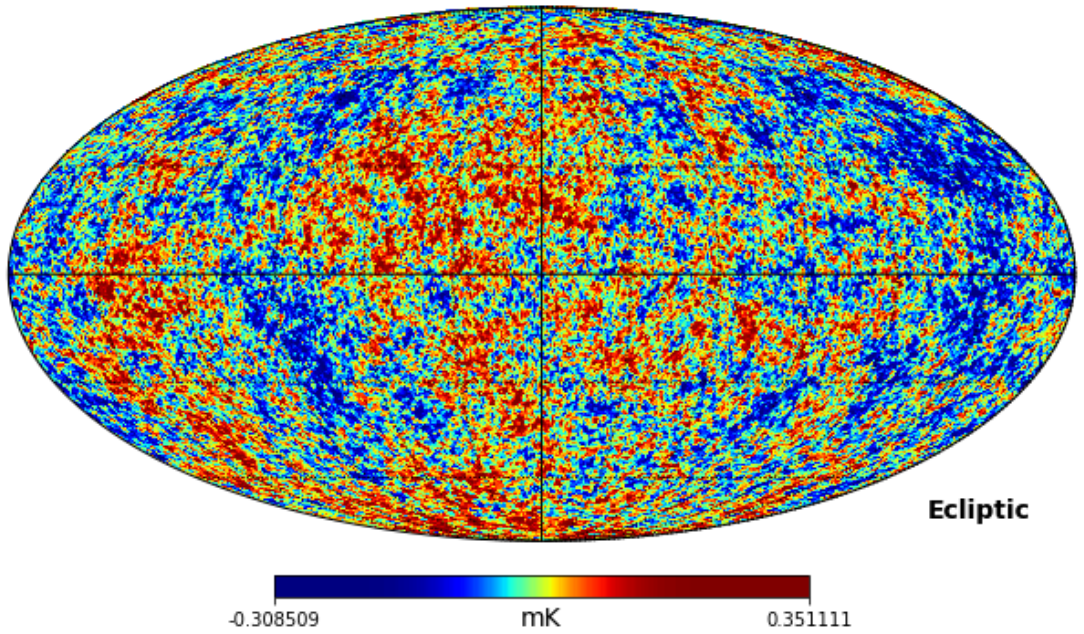
비 정적 데이터는 데이터 분포를 나타내며 시간에 따라 변한다.

실시간 비 정적 데이터는 자주 수집되며 BD 분석에서 까다로운 영역을 제시합니다.

대규모 온라인 실시간 데이터 스트림을 제어 할 수 있도록 DL 알고리즘을 동화하는 것이 중요하다.

https://www.sciencedirect.com/science/article/pii/S0309170815000020

. High-dimensional Data



- 설명

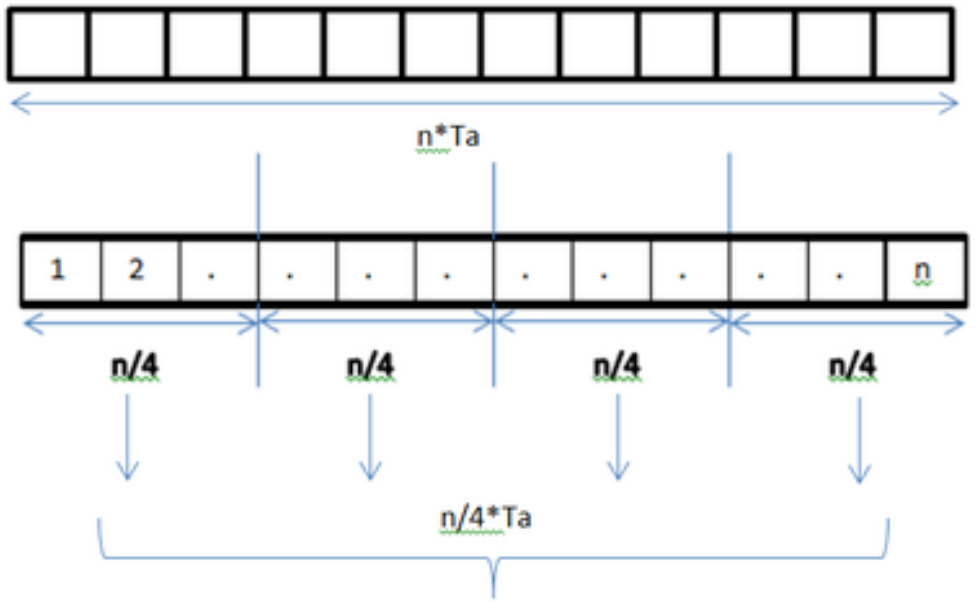
DL 알고리즘과 대용량 BD 사이의 연관성을 수행 할 때, 고차원의 데이터 소스는 입력으로 부터의 학습을 억제한다.

차원 입력을 위해 상당히 확장 된 주변 스택 노이즈 제거 AE를 도입했으며 이는 일반적인 스택 노이즈 제거 자동 인코더보다 체계적

BD Analytics에 DL 알고리즘을 적용하면 주로 사용되지 않는 고차원 데이터가 포함되며 DL 기반 결과의 진화를 보증한다.

https://github.com/lesgourg/class_public/issues/201

. Data parallelism



- 설명

BD에는 대규모 입력, 높은 차원 속성 및 다양한 출력이 포함된다.

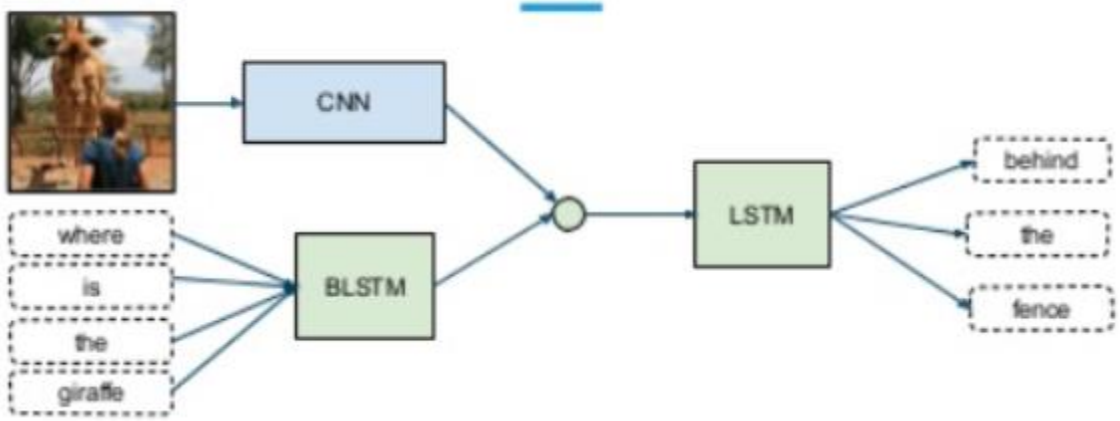
데이터를 실행하기가 매우 복잡해지고 제안 된 모델도 유지하기가 어렵다.

DL 알고리즘은 효과적인 방법을 제공합니다. DL의 인기 있는 SBD 기술은 잘 알려져 있으며 컴퓨터를 통해 병렬화하기가 어렵다.

병렬 구현은 정확도 학습 알고리즘을 줄이지 않고 훈련 데이터 속도를 높이기 위해 GPU 또는 CPU 클러스터를 사용합니다.
-> 최근 TPU처럼 머신러닝을 위한 장치가 개발됨

https://en.wikipedia.org/wiki/Data_parallelism

. Multimodal Data



- 설명

멀티 모달 데이터는 서로 다른 소스에서 제공되는 몇 가지 입력 모드있다.

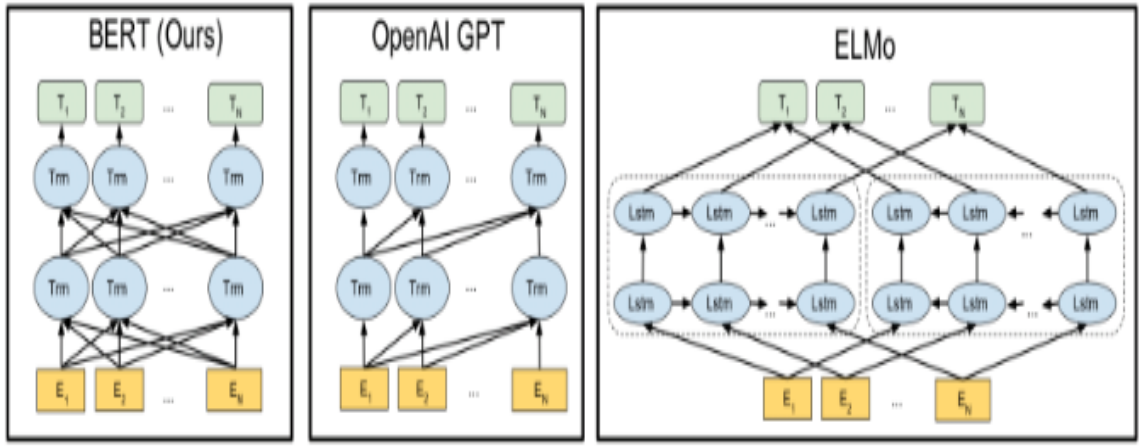
각 소스는 다른 종류의 표현 및 상관 관계 구조를 갖는다.

서로 다른 양식을 통해 저 수준 피처 간의 비선형 관계를 실현하는 것은 어려운 작업

DL은 데이터의 변동 요소를 학습하고 이에 대한 추상 표현을 제공하는 기능으로 인해 두 가지 이상의 데이터 통합에 효과적으로 사용됩니다.

https://towardsdatascience.com/multimodal-deep-learning-ce7d1d994f4

. Large-scale models



- 설명

대규모 DL 모델은 BD와 함께 제공되는 대량의 입력 데이터를 효과적으로 처리한다.

스트리밍 데이터 및 도메인 적응은 BD 분석을 위한 대규모 DL 모델로 효과적으로 해결할 수 있는 다른 BD 문제

좌 사진은 자연어 빅데이터를 분석하기 위한 예시 사진

6. FUTURE WORKOFDEEPLEARNINGINBIGDATA

1. 빅데이터 향 후 연구 방향

. 빅**데이터 향 후** 연구 **방향**

1. 추상화된 데이터들을 학습 하기 위해 DB을 이용하여 특정 패턴을 파악

2. 빅데이터 분석을 위한 DL모델에서 효과적으로 사용하기 위해 얼마나 많은 데이터가 필요한지 정의가 필요

3. 현재 대부분의 빅데이터 DL모델은 특정 도메인에서만 효과적으로 작동 함으로 도메인 적용 연구가 필요하다.

4. 컴퓨터 비전은 레이블이 있는, 없는 데이터가 많은 빅데이터 연구임으로 semi-supervised traning 방법을 통한 패턴을 식별하는 연구가 필요

7. CONCLUSION

1. 결론

결론

1. DL은 분석, 음성 인식, 이미지 처리 및 자연어 처리 등 다양한 영역을 분석한다.

2. 대량의 비 관리데이터에서 복잡한 데이터를 자연스럽게 추출하기 위해 DL을 필요로 한다.

3. 이미지 시맨틱 세그먼테이션, 이미지의 물감 감지, 이미지 분류 등 다양한 영역에서 빅데이터 분석이 잘 수행되고 있지만, 아직 부족한 분석 영역이 많으며 해결이 필요하다.

	분석 필요한 영역
응용 분야	자율 주행 자동차, 건강 관리의 DL, 음성 검색 및 음성 인식 보조 장치
고 차원 데이터	영화에 소리 추가, 자동 기계 번역, 자동 텍스트 생성, 자동 필기 생성, 이미지 인식, 자동 이미지 캡션 생성, 자동 채색, 광고, 지진 예측, 뇌암 감지를위한 신경망, 금융의 신경망, 에너지 시장 가격 예측
DL 알고리즘	데이터 태깅, 향상된 데이터 추상화 공식화, DL 모델의 확장 성, 스트리밍 데이터 분석, BD와 관련된 장애물
기타	시맨틱 인덱싱, 좋은 데이터 표현 선택 기준, 분산 컴퓨팅, 정보 검색 및 도메인 적응.