KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (오프라인)

AI with Real-world Data in Healthcare

고태훈 _ 가톨릭대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 오프라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 : 2월 24일 (토)

시간	강 의 (자연과학대학 28동 101호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:30	**의료빅데이터/인공지능 총론** 김헌성 교수(가톨릭대학교)
14:30-14:45	휴식
14:45-16:15	**의료영상 인공지능의 이해 및 의료영상 레이블링 실습** 백서연 교수(연석대학교)
16:15-16:30	휴식
16:30-18:00	**의료 정보처리 자동화 실습 / 독자적인 어플리케이션 만들기** 김선근 대표(원닥 주식회사), 서사도 조교

시간	강 의 (자연과학대학 28동 102호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:20	**EMR 데이터를 활용한 머신러닝 기반 예후예측:** **Decision Tree-based Models + EMR 샘플 데이터 실습 (MIMIC sample dataset)** 고태훈 교수(가톨릭대학교)
14:20-14:40	휴식
14:40-16:00	**Chest X-ray 영상을 활용한 딥러닝 기반 폐질환 진단:** **Convolutional Neural Network + 의료영상 샘플 데이터 실습 (NIH Chest X-ray14)** 고태훈 교수(가톨릭대학교)
16:00-16:20	휴식
16:20-17:40	**심전도 데이터를 활용한 딥러닝 기반 부정맥 탐지: Recurrent Neural Network +** **Transformer + 심전도 샘플 데이터 실습 (MIT-BIH Arrhythmia Database)** 고태훈 교수(가톨릭대학교)

DAY1 : 2월 26일 (월)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**DNN (이론)** 이상근 교수(고려대학교)
10:50-11:00	휴식
11:00-12:10	**CNN (이론)** 이상근 교수(고려대학교)
12:10-13:40	점심
13:40-15:10	**RNN, ChatGPT, XAI (이론)** 이상근 교수(고려대학교)
15:10-15:20	휴식
15:20-16:50	**CNN/RNN 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)** 이정현 조교, 한성민 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Best practice for single-cell data analysis** 박종은 교수(KAIST)
11:00-11:10	휴식
11:10-12:40	**Practice1: Scanpy basic workflow** 정성민 조교, 고용준 조교
12:40-14:10	점심
14:10-15:30	**Public database, data integration, reference mapping, multiomics** 박종은 교수(KAIST)
15:30-15:40	휴식
15:40-16:50	**Practice2: Advanced single-cell analysis (siVI universe)** 정성민 조교, 고용준 조교

DAY1 : 2월 27일 (화)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**AI-based protein structure prediction** **- Intro to protein structure prediction** **- Early AI-based approaches** **- AlphaFold and RoseTTAFold** 백민경 교수(서울대학교)
10:50-11:00	휴식
11:00-12:10	**단백질 구조 예측 실습** **- ColabFold를 활용한 단백질 구조 및 상호작용 예측** **- Tips &Tricks for better structure modeling** 백민경 교수(서울대학교)
12:10-13:40	점심
13:40-15:10	**AI-based protein design** **- Intro to protein design** **- Protein backbone design using RFdiffusion** **- Protein sequence design using ProteinMPNN** 백민경 교수(서울대학교)
15:10-15:20	휴식
15:20-16:50	**단백질 디자인 실습** **- RFdiffusion 및 ProteinMPNN의 활용법 실습** 백민경 교수(서울대학교)

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Single-cell biology** 최정민 교수(고려대학교)
11:00-11:10	휴식
11:10-12:40	**i. Unsupervised Spatial transcriptome analysis** **ii. Tumor Boundary Determination in Spatial Transcriptomics** 유광민 조교, 이문영 조교
12:40-14:10	점심
14:10-15:30	**i. Deconvolution Analysis Using Single-cell RNA Sequencing and Spatial Transcriptomics** **ii. Cell-Cell Interaction Analysis in Spatial Transcriptomics** 김지현 조교, 최승지 조교
15:30-15:40	휴식
15:40-16:50	**i. Open Chromatin Region Analysis and Biological Interpretation of Using scATAC-seq Dataset** **ii. Construction of Gene Regulatory Networks Based on Integrated Analysis of scATAC-seq and scRNA-seq Datasets** 천하림 조교, 이호진 조교

DAY1 : 2월 28일 (수)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Transformers (이론)** 전민지 교수 (고려대학교)
11:00-11:10	휴식
11:10-12:40	**Introduction to Transformers (실습)** 봉현수 조교, 임우택 조교
12:40-14:10	점심
14:10-15:40	**Deep learning in Bioinformatics** 노미나 교수(한양대학교)
15:40-15:50	휴식
15:50-17:20	**Deep learning model을 이용한 실습** 박예솔 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**마이크로바이옴 기본 이론** 이선재 교수(GIST)
10:50-11:00	휴식
11:00-12:10	**16S rRNA amplicon seq. - DADA2** 조준우 조교, 백재우 조교
12:10-13:40	점심
13:40-14:40	**최신 메타지놈 분석 기법의 현황** 이선재 교수(GIST)
14:40-14:50	휴식
14:50-16:50	**Shotgun metagenome 분석 (Linux)** 조준우 조교, 백재우 조교

DAY1 : 2월 29일 (목)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**화학정보학 기초(Cheminformatics) / 약물특성 및 약물다움(druglikeness)** **Molecular Notations &Descriptors / AI 신약개발을 위한 Databases** **AI 신약개발을 위한 Programming 기초** 김동섭 교수(KAIST)
10:50-11:00	휴식
11:00-12:10	**Google Colab에 RDKit 설치 / 화합물 정보 읽기 실습** **Bioactivity database 검색 및 정보 읽기 실습** **Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습** 정수재 조교, 나민주 조교
12:10-13:40	점심
13:40-15:10	**AI 신약개발을 위한 기계학습법 기초 / QSAR 모델링 기초 / AI 신약개발을 위한 딥러닝 모델** **Virtual screening (ligand-based, structure-based) 및 de novo design** 김동섭 교수(KAIST)
15:10-15:20	휴식
15:20-16:50	**QSAR modeling 전체 과정 실습/ 화합물의 Bioactivity 예측 모델 개발** **Virtual screening 과정을 통한 신약후보물질 발굴 실습** 정수재 조교, 나민주 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Single cell multiomics 이론 / Gene regulatory network 이론** 김준일 교수(숭실대학교)
11:00-11:10	휴식
11:10-12:40	**Seurat/Signac, ArchR, TENET+ 실습** 김현규 조교, 정회빈 조교
12:40-14:10	점심
14:10-15:40	**롱리드 시퀀싱 소개 및 유전체 조립 실습** 김준 교수(충남대학교)
15:40-15:50	휴식
15:50-17:20	**변이 분석 및 시각화 실습** 김준 교수(충남대학교)

강의개요

AI with Real-world Data in Healthcare

실제 의료 현장에서는 다양한 종류의 데이터가 생성되고, 이에 따라 적용할 수 있는 AI 기술에도 조금씩 차이가 있다. 본 강의에서는 실제 의료 분야에서 AI 연구가 활발히 진행되고 있는 전자의무기록 (Electronic Medical Record, EMR) 데이터, 의료영상 데이터, 그리고 실시간 생체신호 데이터에 대한 머신러닝 및 딥러닝 알고리듬을 빠르게 배운 후 이에 대한 실습을 진행하고자 한다. 사용하는 데이터는 public access가 가능한 데이터를 사용할 예정이며, 모든 실습은 Python 프로그래밍 언어로 진행한다. 실습은 Google Colaboratory를 사용할 예정이므로 인터넷 접속이 가능한 노트북을 지참하면 누구나 참여가 가능하다.

강의는 다음의 내용을 포함한다:

- EMR 데이터를 위한 Decision tree-based models: Random Forest and XGBoost
- 의료영상 데이터를 위한 Convolutional Neural Network (CNN): 사전학습된 모델의 파인튜닝
- 실시간 생체신호 데이터를 위한 Long Short-Term Memory (LSTM), Transformer 모델

* 참고강의교재: 교수자의 강의자료

* 교육생준비물: 노트북 (노트북 사양은 중요하지 않으며, Wi-fi 접속 필요)

* 강의 난이도: 중급/고급

* 강의: 고태훈 교수 (가톨릭대학교 의과대학 의료정보학교실) / 이강혁 조교

Curriculum Vitae

Speaker Name: Taehoon Ko, Ph.D.

▶ **Personal Info**

Name	Taehoon Ko
Title	Assistant Professor
Affiliation	Department of Medical Informatics, College of Medicine, The Catholic University of Korea

▶ **Contact Information**

Address	222, Banpo-daero, Seocho-gu, Seoul, 06591
Email	thko@catholic.ac.kr
Phone Number	010-3494-5445

Research Interest

Machine learning and artificial intelligence in healthcare

Educational Experience

2008	B.S. in Industrial Engineering, Seoul National University, Korea
2010	M.S. in Industrial Engineering, Seoul National University, Korea
2017	Ph.D. in Industrial Engineering, Seoul National University, Korea

Professional Experience

2017-2020	Research Assistant Professor, Office of Hospital Information, Seoul National University Hospital, Korea
2020-2022	Research Assistant Professor, Department of Medical Informatics, College of Medicine, The Catholic University of Korea
2022-Current	Assistant Professor, Department of Medical Informatics, College of Medicine, The Catholic University of Korea

Selected Publications (5 maximum)

1. Joo, M. W., Ko, T., Kim, M. S., Lee, Y. S., Shin, S. H., Chung, Y. G., & Lee, H. K. (2022). Development and validation of a convolutional neural network model to predict a pathologic fracture in the proximal femur using abdomen and pelvis CT images of patients with advanced cancer. Clinical Orthopaedics and Related Research®, 10-1097.
2. Oh, G.C., Ko, T., Kim, J.H., Lee, M.H., Choi, S.W., Bae, Y.S., Kim, K.H. & Lee, H.Y. (2022). Estimation of low-density lipoprotein cholesterol levels using machine learning. International Journal of Cardiology, 352, 144-149.
3. Kim, H. M., Ko, T., Choi, I. Y., & Myong, J. P. (2022). Asbestosis diagnosis algorithm combining the lung segmentation method and deep learning model in computed tomography image. International Journal of Medical Informatics, 158, 104667.
4. An, Y., Lee, S., Jung, S., Park, H., Song, Y., & Ko, T. (2021). Protect: Privacy-preserving contact tracing for COVID-19 with homomorphic encryption. J. Med. Internet Res., 23(7).
5. Jo, C., Ko, S., Shin, W. C., Han, H. S., Lee, M. C., Ko, T., & Ro, D. H. (2020). Transfusion after total knee arthroplasty can be predicted using the machine learning algorithm. Knee Surgery, Sports Traumatology, Arthroscopy, 28, 1757-1764.

KSBI | 한국생명정보학회

KSBi-BIML 2024

AI with Real-world Data in Healthcare

가톨릭대학교 의과대학 의료정보학교실

고태훈 (thko@catholic.ac.kr)

1

워크샵의 목표

- 의료 데이터에 대한 머신러닝, 딥러닝의 적용 “체험”
  - 완전 초보자 프로젝트보다는 어려움
  - 향후 의료 데이터에 대한 AI 연구를 직접 하실 분에게는 속성교육 및 방향성 제시
  - 직접 하지 않더라도 AI 연구자들과 협업 시 원활한 소통에 도움

2

시 간	강 의 내 용	연 자
12:30-12:50(20)	등 록	
12:50-13:00(10)	공지사항 전달	
13:00–14:20(80)	EMR 데이터를 활용한 머신러닝 기반 예후예측: Decision Tree-based Models + EMR 샘플 데이터 실습 (MIMIC sample dataset)	고태훈 교수
14:20–14:40(20)	휴 식	
14:40-16:00(80)	Chest X-ray 영상을 활용한 딥러닝 기반 폐질환 진단: Convolutional Neural Network + 의료영상 샘플 데이터 실습 (NIH Chest X-ray14)	고태훈 교수
16:00-16:20(20)	휴 식	
16:20-17:40(80)	심전도 데이터를 활용한 딥러닝 기반 부정맥 탐지: Recurrent Neural Network + Transformer + 심전도 샘플 데이터 실습 (MIT-BIH Arrhythmia Database)	고태훈 교수

3

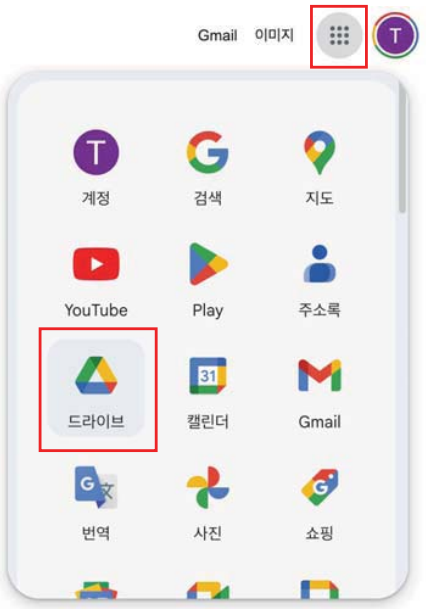
실습환경: Google Colaboratory (Colab)

4

Google Colab 환경 만들기

1. 웹 브라우저로 구글 (https://www.google.com) 접속

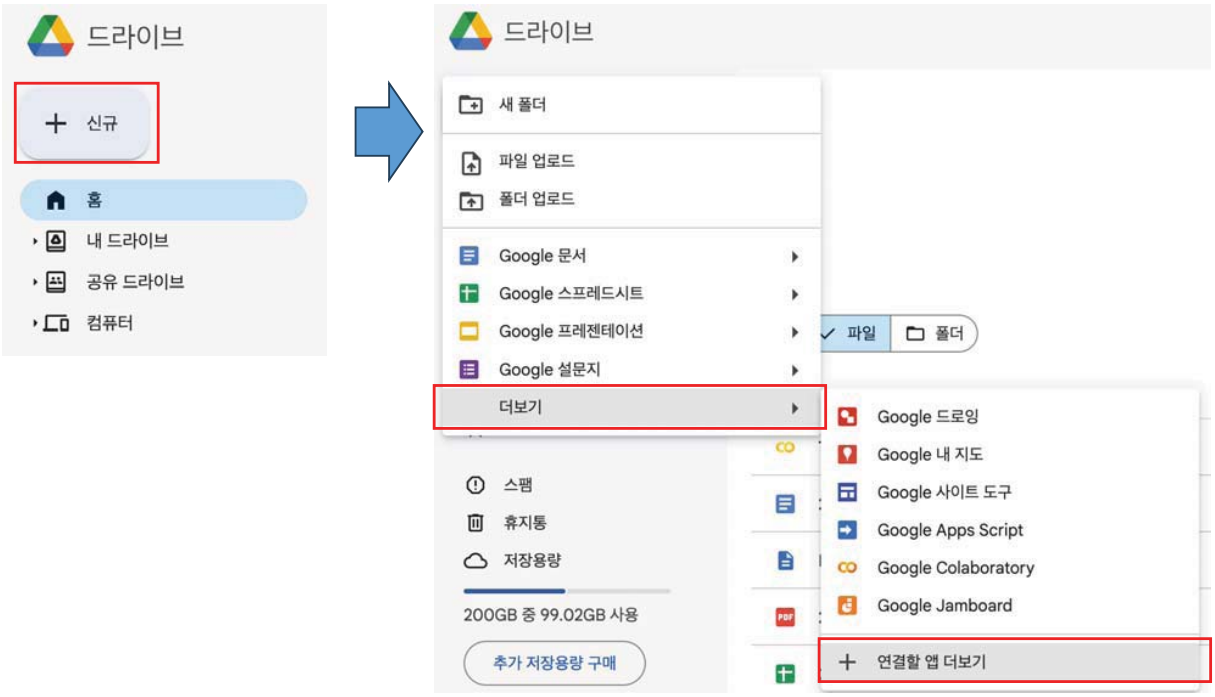
2. 구글 로그인 후, 구글 드라이브 접속



5

Google Colab 환경 만들기

3. 왼쪽 메뉴에서 [+ 신규] 버튼 누른 후, 메뉴에서 [더보기] → [+ 연결할 앱 더보기] 클릭

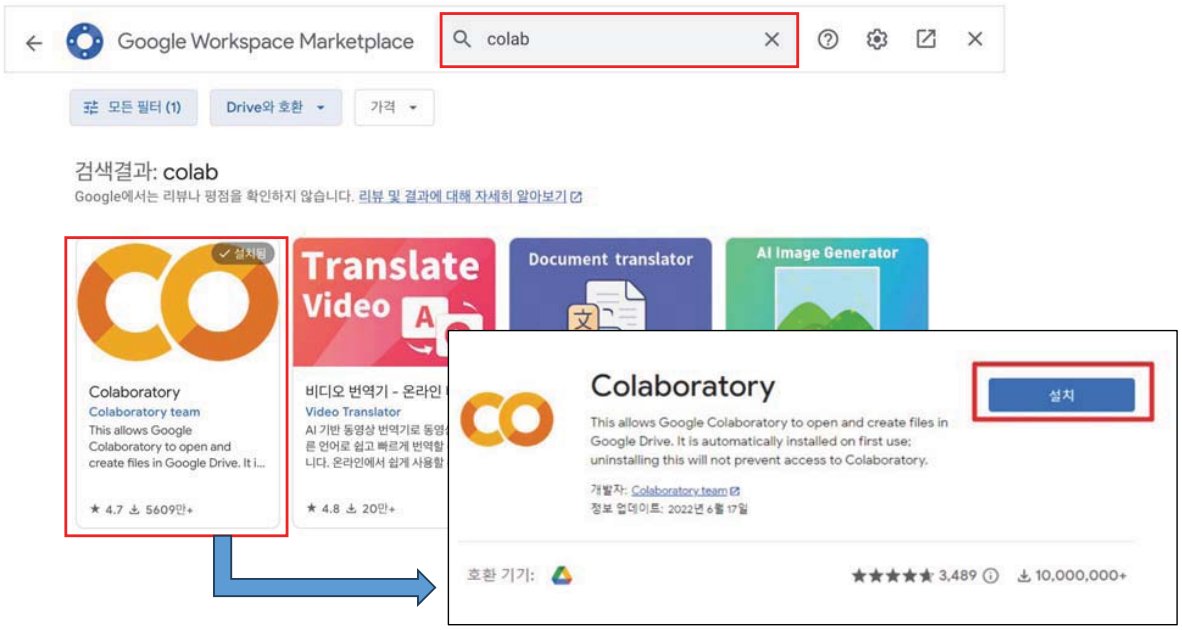


6

Google Colab 환경 만들기

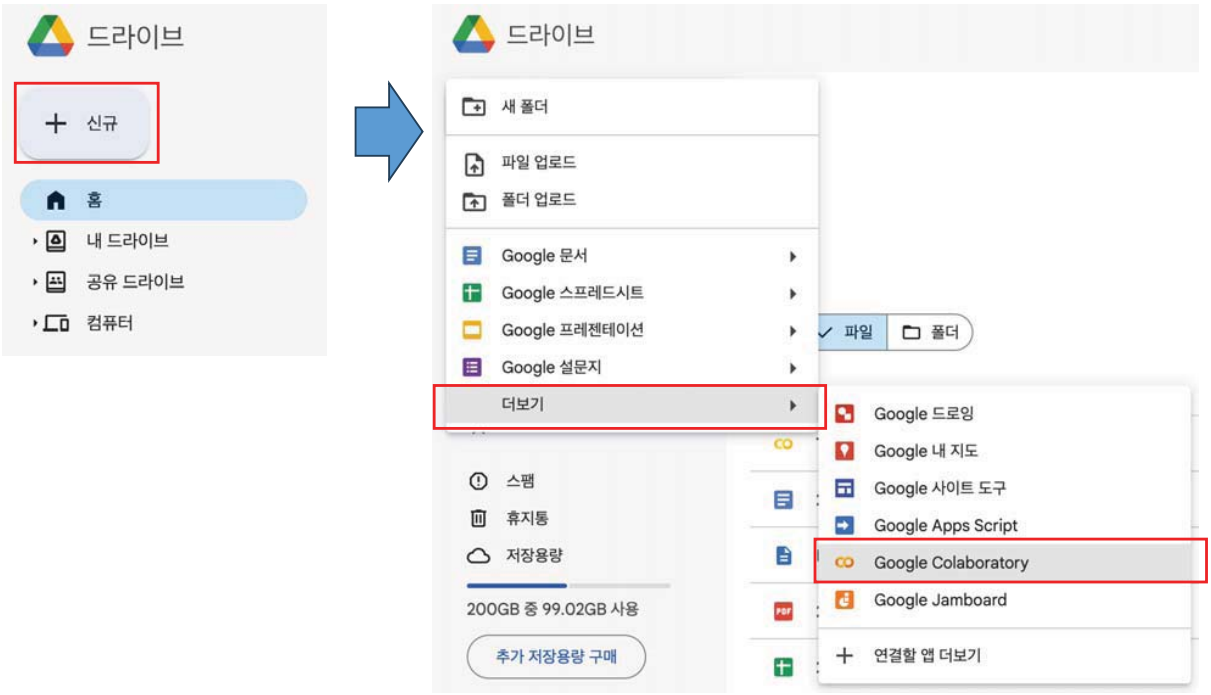
4. Google Workspace Marketplace에서 “colab” 검색.

5. 검색결과 중 “Colaboratory”를 설치



Google Colab 환경 만들기

6. 왼쪽 메뉴에서 [+ 신규] 버튼 누른 후, 메뉴에서 [더보기] → [Google Colaboratory] 클릭



8

Google Colab 환경 만들기

7. 현재 구글 드라이브에 Untitled0.ipynb 라는 파일이 생성됨.
   여기에서 바로 코드를 작성할 수 있음

(2023년 12월부터 프롬프트를 사용하여 자동 코드 생성이 가능함)

8. 이제부터는 구글 드라이브에서 바로 *.ipynb 를 더블클릭해서 실행하거나 Google Colaboratory를 연결할 앱으로 선택하면 위와 같은 창을 볼 수 있음

9

Google Colab 환경 만들기

7. 현재 구글 드라이브에 Untitled0.ipynb 라는 파일이 생성됨.
   여기에서 바로 코드를 작성할 수 있음

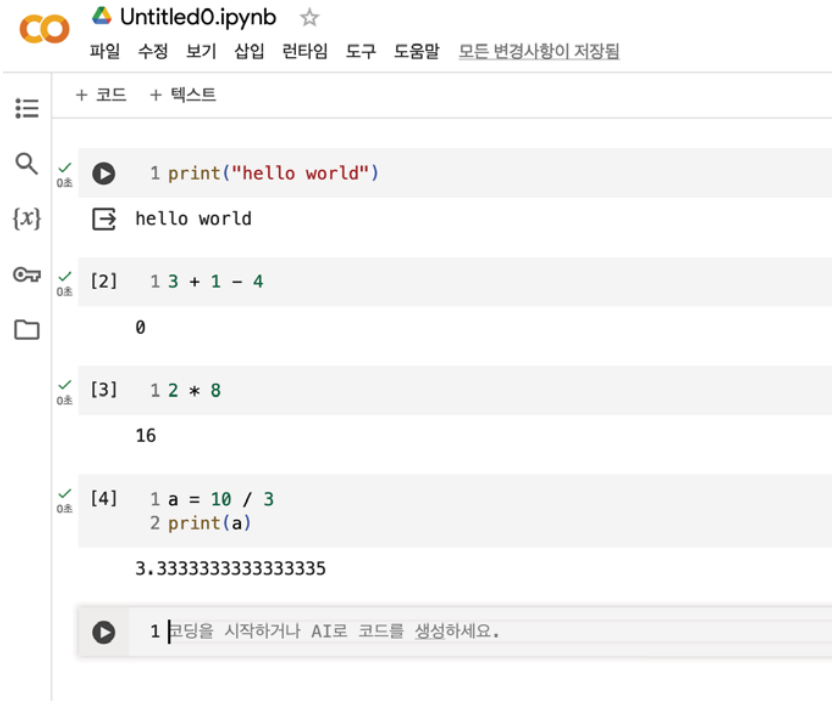
(2023년 12월부터 프롬프트를 사용하여 자동 코드 생성이 가능함)

8. 이제부터는 구글 드라이브에서 바로 *.ipynb 를 더블클릭해서 실행하거나 Google Colaboratory를 연결할 앱으로 선택하면 위와 같은 창을 볼 수 있음

10

Google Colab 환경 만들기

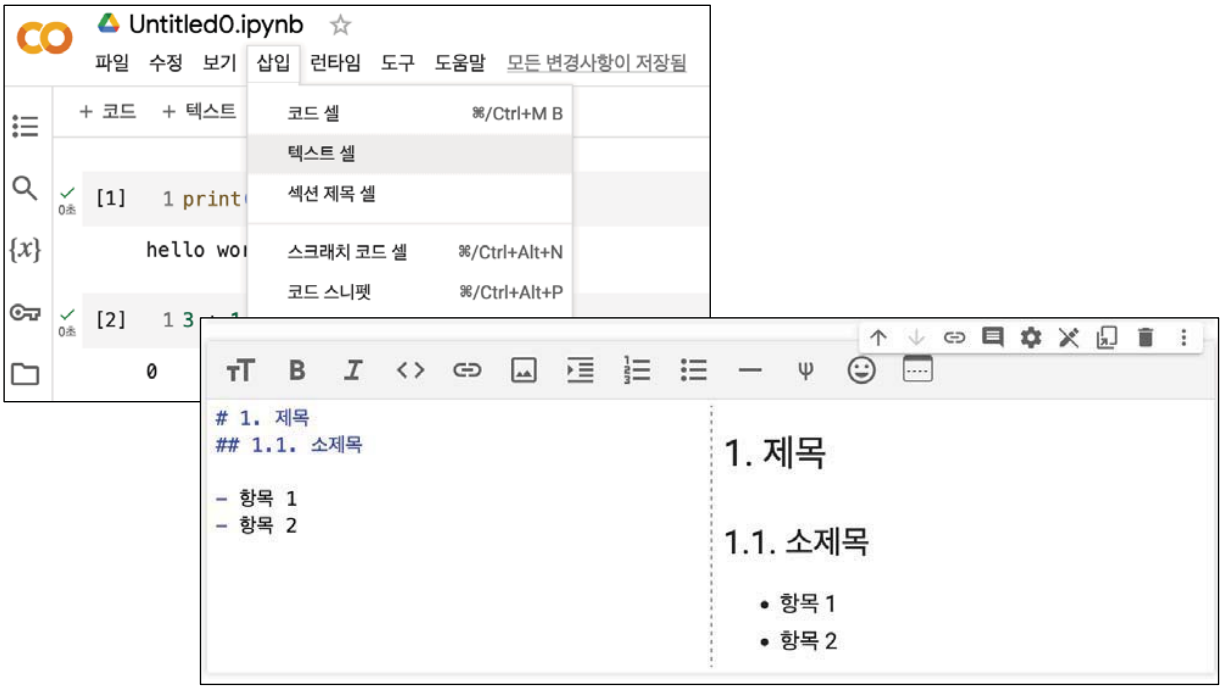
9. 코드에 명령을 입력하고 실행시켜보기
   - 실행방법 1: 해당 cell에 나타나는 play button 클릭
   - 실행방법 2: Shift+Enter 또는 Ctrl+Enter



11

Google Colab 환경 만들기

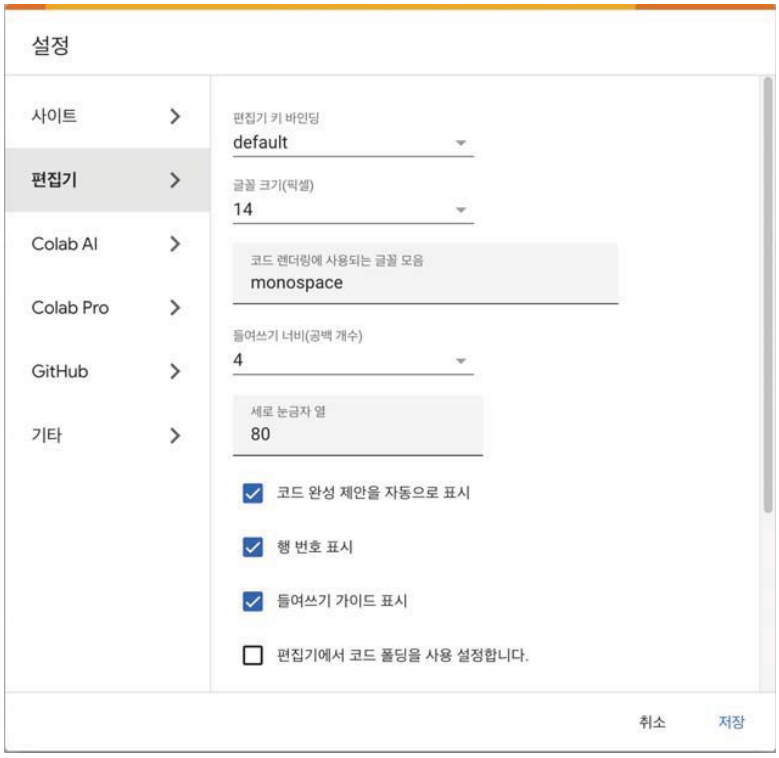
10. 텍스트를 입력하는 셀도 추가 가능
   - Markdown 이라는 방법으로 작성함 (Markdown 작성법 예시: link)
   - 텍스트 셀 삽입 방법 1: 메뉴의 [삽입] → [텍스트 셀]



12

Google Colab 환경 만들기

11. Colab 편집기 주요 설정 (메뉴에서 [도구] → [설정] → [편집기])



13

Google Colab 환경 만들기

12. Colab 런타임 관리 (메뉴의 [런타임])

- **세션 다시 시작**: 현재의 컴퓨팅 환경은 유지된 채, ipynb 이 실행된 것을 초기화
- **런타임 연결 해제 / 유형 변경**: 현재의 컴퓨팅 환경을 새로 바꿈

14

Google Colab 사용 유의점

- 장점
  - Colab은 파이썬 환경을 사용자가 직접 만들지 않아도 됨
  - 내 Google Drive에 데이터셋을 넣고 원하는 분석을 할 수 있음
  - Free tier는 규모가 작은 연구를 돌리는 데에도 충분하며, 복잡한 딥러닝 연구의 proof-of-concept 수준은 충분히 구현할 수 있음
- 유의점
  - 세션, 런타임 구동 시간, 횟수가 제한적
    (Tier를 높이면 구동 시간, 횟수를 더 늘릴 수 있음)
  - GPU 런타임으로 돌림에도 불구하고, GPU를 활용하지 않는 경우 페널티가 발생

15

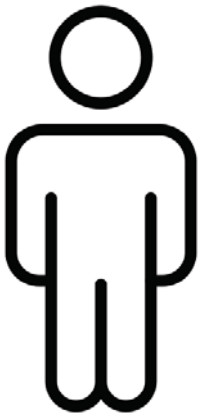
Introduction

16

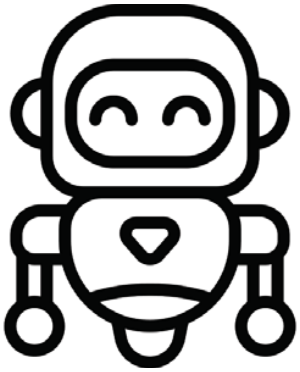
What is Machine Learning?

- Machine learning
  - branch of artificial intelligence (AI)
  - techniques that enable computers to learn from data

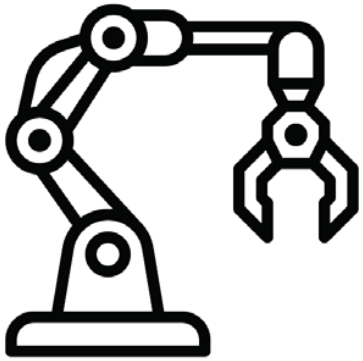
Learn from experience



Learn from *data*



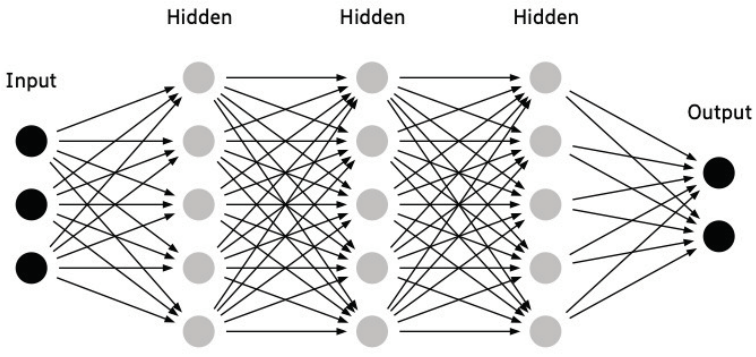
Follow the instruction



17

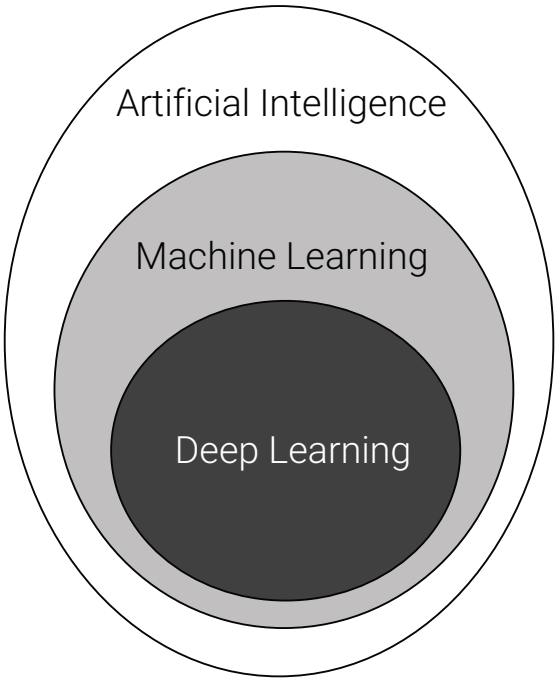
What is Deep Leaning?

- Deep Learning
  - branch of machine learning
  - an advanced form of 'artificial neural network' inspired by the human brain



18

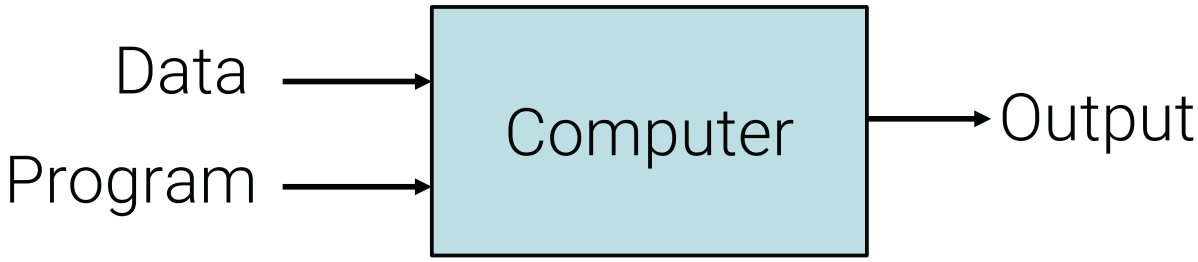
AI, ML and DL



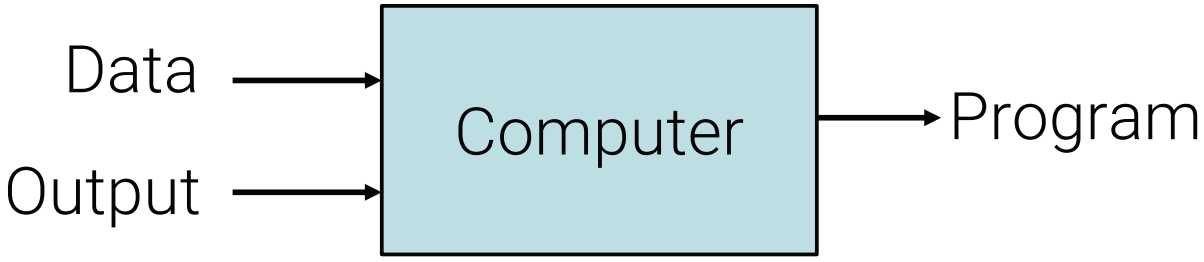
- **AI is the broadest concept.**
  It encompasses all attempts to create intelligent machines.
- **ML is a subfield of AI.**
  It focuses on using algorithms to learn from data and improve performance over time. This is one method to achieve AI.
- **DL is a subfield of ML.**
  It uses artificial neural networks with multiple layers to learn complex patterns from data.

19

Classic Programming



Machine Learning



20

Machine Learning and Statistics

- Similarities
  - Both rely on data
  - Use mathematical models
- Differences
  - ML is about building powerful predictive models, often at the cost of interpretability.
  - Statistics is about understanding the underlying structures and relationships within data, focusing on interpretable results, and generalization to unseen data.

21

Machine Learning in Healthcare

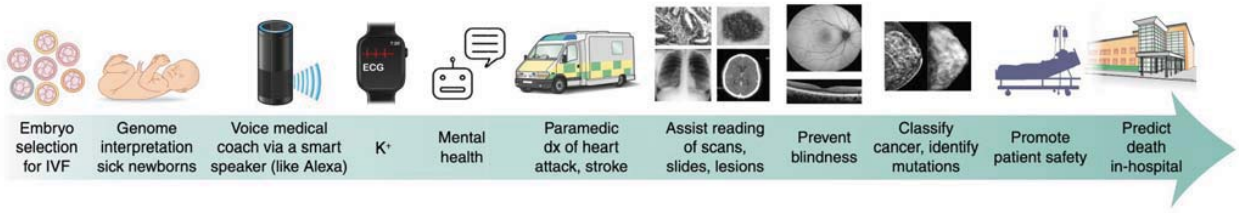


Fig. 2 | Examples of AI applications across the human lifespan. dx, diagnosis; IVF, in vitro fertilization K+, potassium blood level. Credit: Debbie Maizels/ Springer Nature

Table 3 | Selected reports of machine- and deep-learning algorithms to predict clinical outcomes and related parameters

Prediction	*n*	AUC	Publication (Reference number)
In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis	216,221	0.93*0.75+0.85#	Rajkomar et al.[96]
All-cause 3–12 month mortality	221,284	0.93^	Avati et al.[91]
Readmission	1,068	0.78	Shameer et al.[106]
Sepsis	230,936	0.67	Horng et al.[102]
Septic shock	16,234	0.83	Henry et al.[103]
Severe sepsis	203,000	0.85@	Culliton et al.[104]
Clostridium difficile infection	256,732	0.82++	Oh et al.[93]
Developing diseases	704,587	range	Miotto et al.[97]
Diagnosis	18,590	0.96	Yang et al.[90]
Dementia	76,367	0.91	Cleret de Langavant et al.[92]
Alzheimer's Disease (+ amyloid imaging)	273	0.91	Mathotaarachchi et al.[98]
Mortality after cancer chemotherapy	26,946	0.94	Elfiky et al.[95]
Disease onset for 133 conditions	298,000	range	Razavian et al.[105]
Suicide	5,543	0.84	Walsh et al.[86]
Delirium	18,223	0.68	Wong et al.[100]

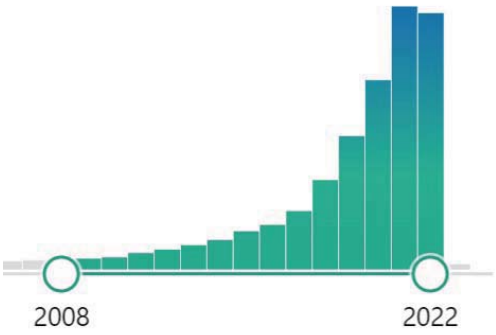
LOS, length of stay; *n*, number of patients (training+validation datasets). For AUC values: *, in-hospital mortality; +, unplanned readmission; #, prolonged LOS; ^, all patients; @, structured+unstructured data; ++, for University of Michigan site.

Source: High-performance medicine: the convergence of human and artificial intelligence, Eric Topol, Nature Medicine, 2019.

22

의학 분야에서 의료 인공지능 연구의 급증

- 의료 인공지능
  = (의학 통계와 마찬가지로) 하나의 연구 수단



PubMed Search query:
"machine learning"

Year	Count
2022	24765
2021	25425
2020	18163
2019	12647
2018	8328
2017	5278
2016	3915
2015	3286
2014	2418
2013	1923
2012	1487
2011	1139
2010	713
2009	595
2008	499

PubMed Search query:
"deep learning"

Year	Count
2022	15978
2021	14567
2020	9288
2019	5598
2018	3082
2017	1352
2016	642
2015	384
2014	265
2013	207
2012	146
2011	118
2010	107
2009	106
2008	105

23

Agenda

- Tree-based Models
  - 테이블 형태의 데이터셋에서 가장 좋은 성능을 보이고 있는 ML 모델들
  - 단일 의사결정나무 (Decision Tree) 를 응용하여 만든
    랜덤 포레스트 (Random Forest) 와 그레디언트 부스팅 머신 (Gradient Boosting Machine, especially XGBoost)
- Convolutional Neural Network (CNN)
  - 이미지, 비디오 데이터의 분류, 분할 등에 널리 사용
  - 대규모 이미지 데이터셋으로 사전학습된 CNN 모델 (pretrained CNN) 을 나의 데이터셋을 이용해 미세조정(fine-tuning)하는 전이학습(transfer learning)
- Recurrent Neural Network(RNN)
  - 실시간 신호, 자연어처리 등에 널리 사용

24

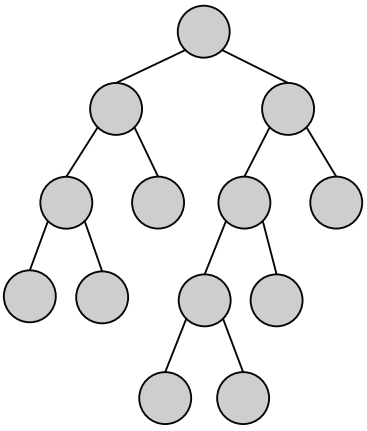
Tree-based Models

- Decision Tree:

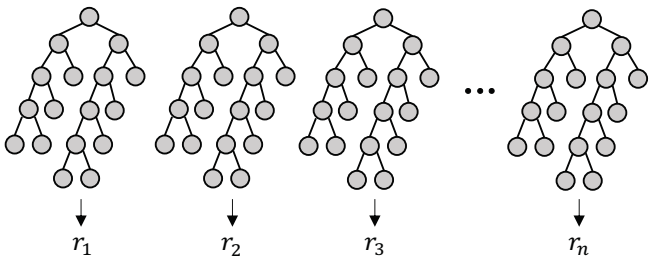
  Classification And Regression Tree (CART)
- Random Forest
- Gradient Boosting Machine and XGBoost

Tree-based Models

Classification And Regression Tree (CART)

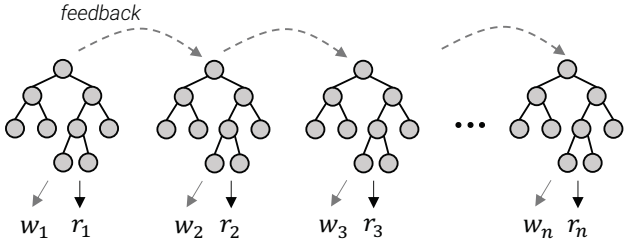


Random Forest (RF)



Combined result: $\frac{1}{n}\sum_{i=1}^{n} r_i$

Gradient Boosting Machine (GBM)



Combined result: $\sum_{i=1}^{n} w_i r_i$

Tree-based Models

- **테이블 형태의 데이터에서 가장 좋은 성능을 보임**
  - 2010년대 초중반에는 Random Forest (RF), 2017년 이후부터는 XGBoost, LightGBM, CatBoost 등의 Gradient Boosting Machine (GBM) 이 널리 사용되고 있음
  - 이유
    - Decision Tree 학습 시 입력변수(독립변수)에 대한 평가와 선택이 끊임없이 이루어짐
    - Decision Tree 가 매우 유연함
      ➔ 여러 ML 모델들을 학습하는 앙상블 기술에 최적임
- **ML 모델의 결과에 대한 해석이 (어느 정도) 가능함**
  - 변수의 중요도 (Feature Importance)
  - SHapley Additive exPlanations (SHAP) 등
- **데이터의 다양한 자료형** (Real, Categorical **등) 에 모두 대응 가능**
- **결측치가 있어도 학습 가능**

27

Tree-based Models

- Decision Tree:
  Classification And Regression Tree (CART)
- Random Forest
- Gradient Boosting Machine and XGBoost

28

의사결정나무 예시 : 포유류 분류

- 여러 동물들에 대한 정보를 조사하여 다음과 같은 데이터가 있다고 가정해보자.

ID	Body Temperature	Gives Birth?	Weight(kg)	…	Mammal?
1	Warm	Yes	65	…	Mammal
2	Warm	No	0.32	…	Non-mammal
3	Cold	No	0.14	…	Non-mammal
4	Warm	Yes	3,000	…	Mammal
5	Cold	No	127	…	Non-mammal
6	Cold	No	0.35	…	Non-mammal
7	Warm	Yes	1.5	…	Mammal
…	…	…	…	…	…

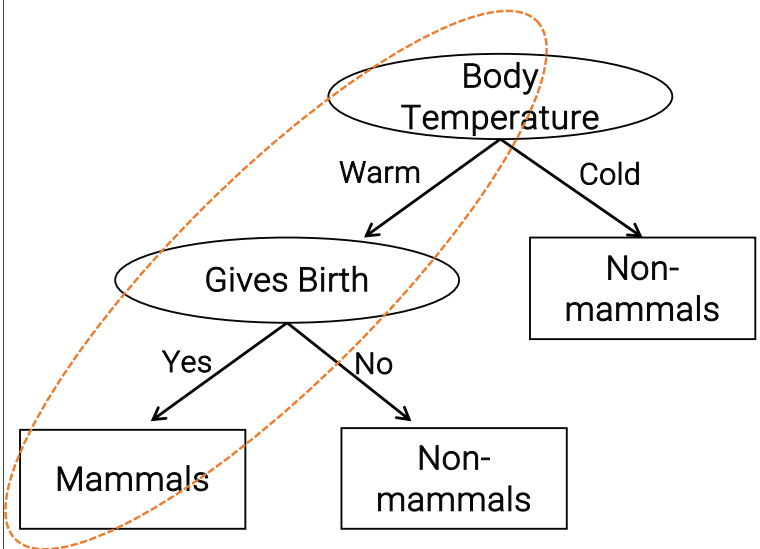
Body Temperature = Warm ➔ 온혈동물(혹은 정온동물)
Body Temperature = Cold ➔ 냉혈동물(혹은 변온동물)

Gives Birth = Yes ➔ 태생
Gives Birth = No ➔ 난생

29

의사결정나무 예시 : 포유류 분류

- 앞의 데이터를 학습하여 도출된 의사결정나무 모델은 다음과 같다.



▶ **"체온"과 "새끼를 낳는 방법"이 포유류를 분류하는 데에 있어 중요한 변수라는 것을 알 수 있다.**

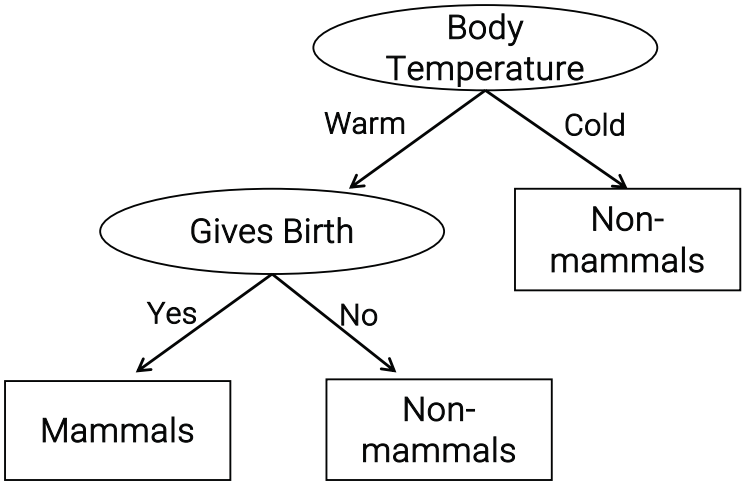
▶ **의사결정나무를 통해 포유류으로 분류하는 규칙을 구할 수 있다.**

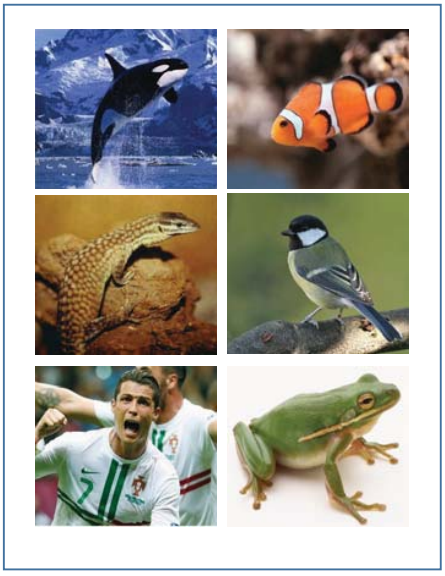
- **온혈동물이고 태생이면 포유류이다.**
- *If* Body Temperature = Warm *and* Gives Birth = Yes, *then* Mammals.

30

의사결정나무 예시 : 포유류 분류

- 새로운 데이터를 의사결정나무 모델을 이용하여 분류해보자.



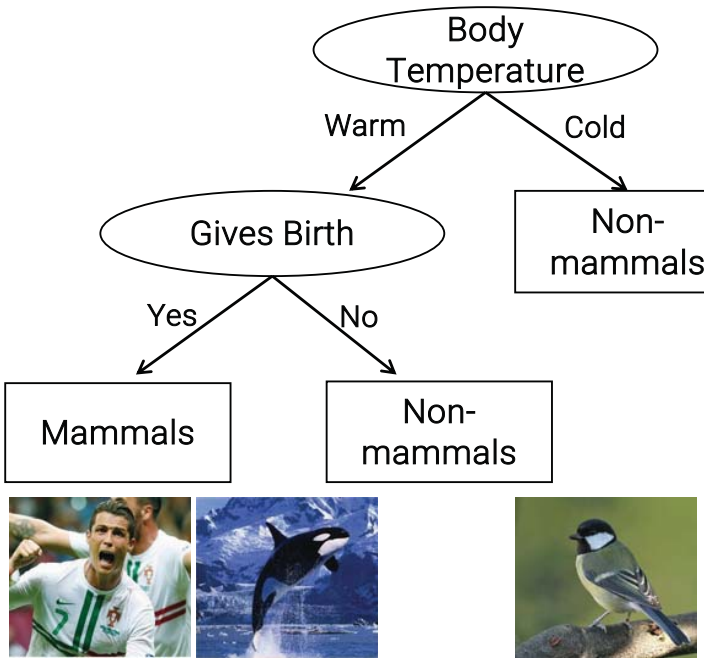


새로운 데이터

31

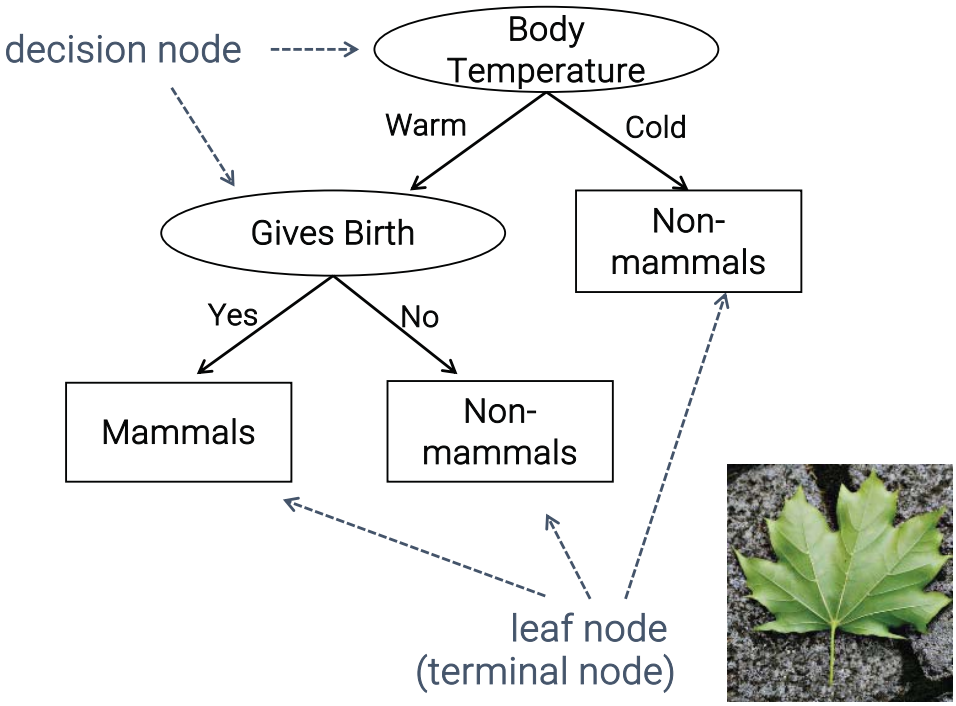
의사결정나무 예시 : 포유류 분류

- 분류 완료



32

의사결정나무의 구성



33

Decision Tree Example

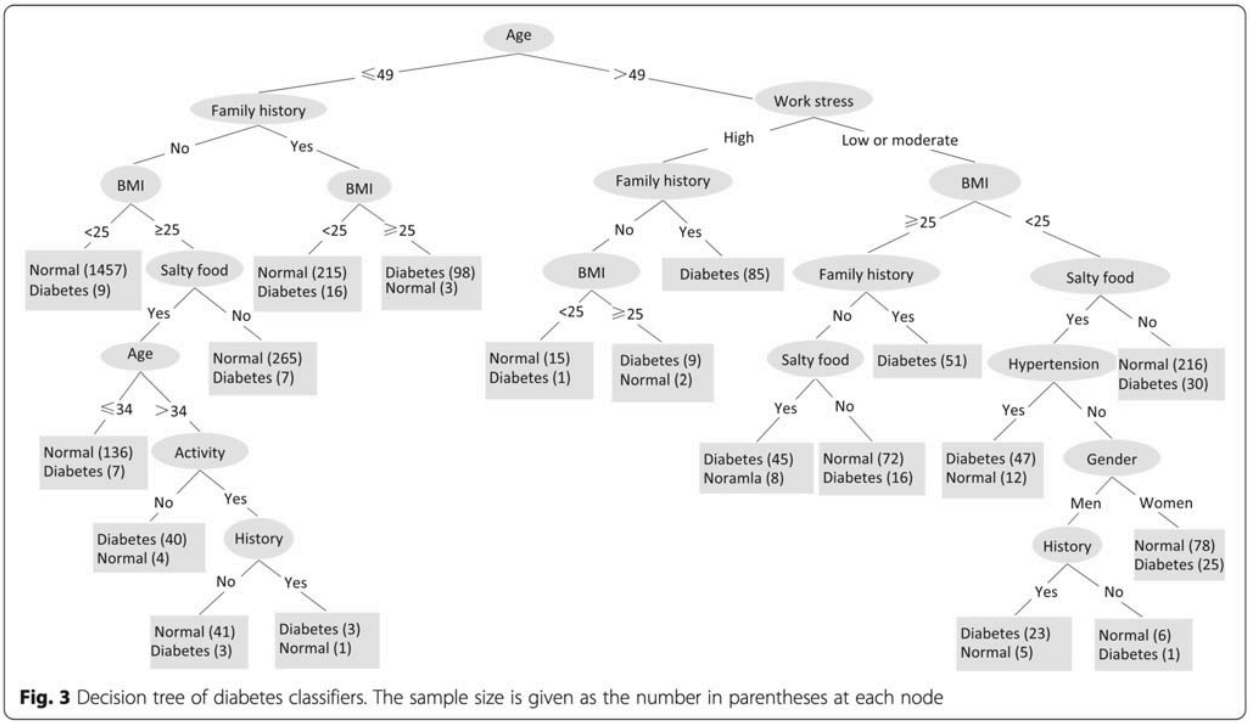


Fig. 3 Decision tree of diabetes classifiers. The sample size is given as the number in parentheses at each node

Source: Accurate and rapid screening model for potential diabetes mellitus, Pei et al., BMC Medical Informatics and Decision Making, 2019

34

Tree-based Models

- Decision Tree:

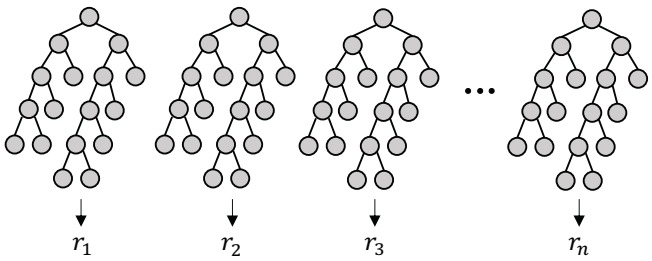
  Classification And Regression Tree (CART)
- Random Forest
- Gradient Boosting Machine and XGBoost

35

Tree-based Models

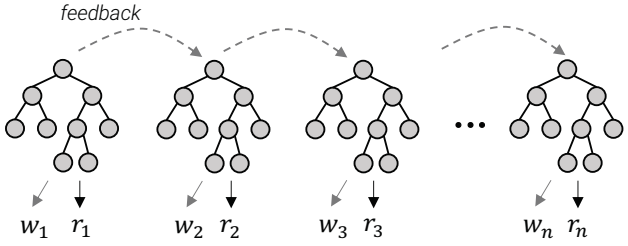
- In RF, trees are trained in
  - fully,
  - parallelly, and
  - independently of each other
- In GBM, trees are trained in
  - not fully,
  - sequentially, and
  - dependently (the next tree is trained in a way that reduces the error of previous trees)

Random Forest (RF)



Combined result: $\frac{1}{n}\sum_{i=1}^{n} r_i$

Gradient Boosting Machine (GBM)

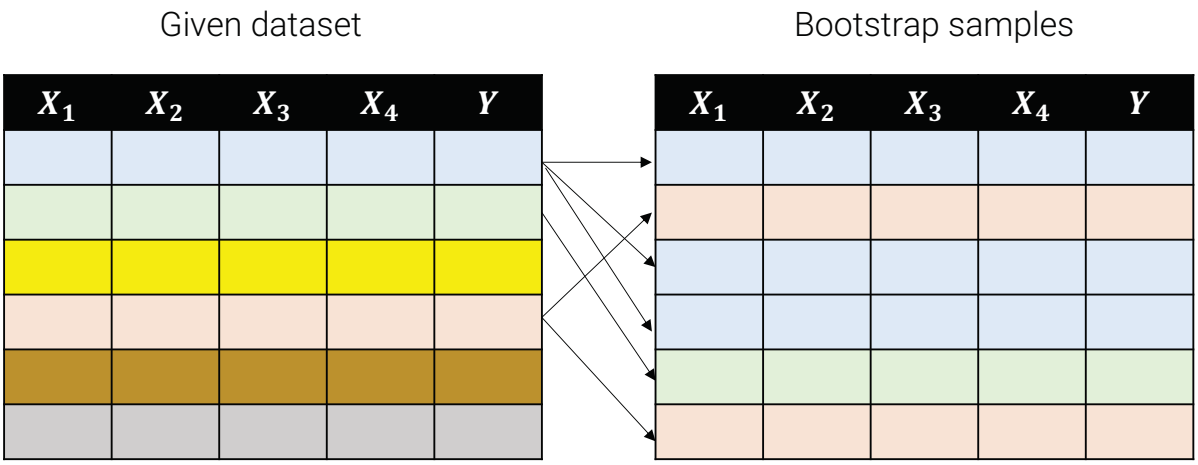


Combined result: $\sum_{i=1}^{n} w_i r_i$

36

Random forest: 2 randomizations

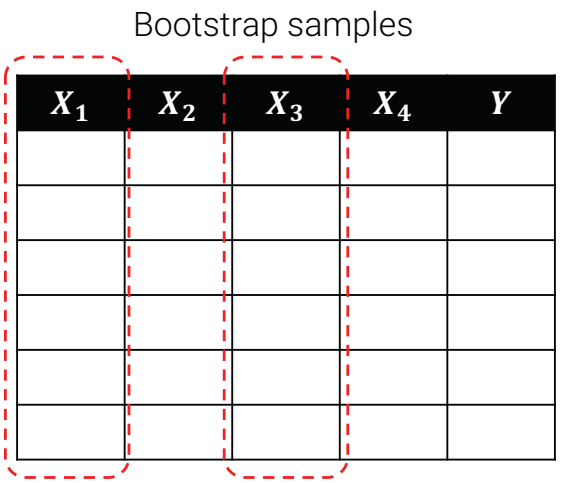
- 2 randomizations are applied.
  - 1st randomization: bagging

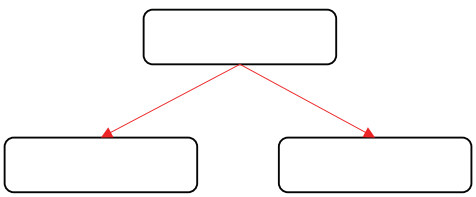


37

Random forest: 2 randomizations

- 2 randomizations are applied.
  - 1st randomization: bagging
  - 2nd randomization: input variable subsets chosen randomly at each split





select *k* input variables randomly.

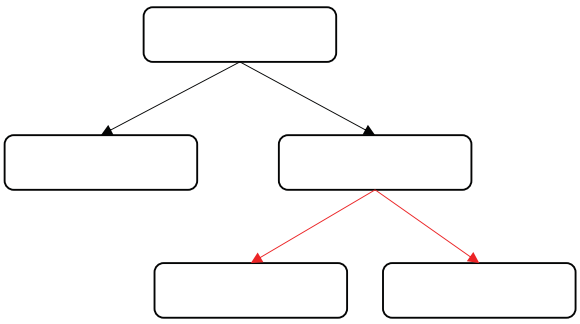
38

Random forest: 2 randomizations

- 2 randomizations are applied.
  - 1st randomization: bagging
  - 2nd randomization: input variable subsets chosen randomly at each split

Bootstrap samples

X_1	X_2	X_3	X_4	Y



select *k* input variables randomly.

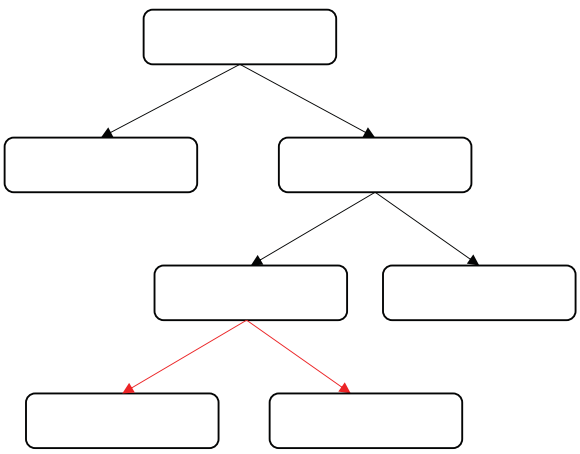
39

Random forest: 2 randomizations

- 2 randomizations are applied.
  - 1st randomization: bagging
  - 2nd randomization: input variable subsets chosen randomly at each split

Bootstrap samples

X_1	X_2	X_3	X_4	Y



select *k* input variables randomly.

40

Tree-based Models

- Decision Tree:

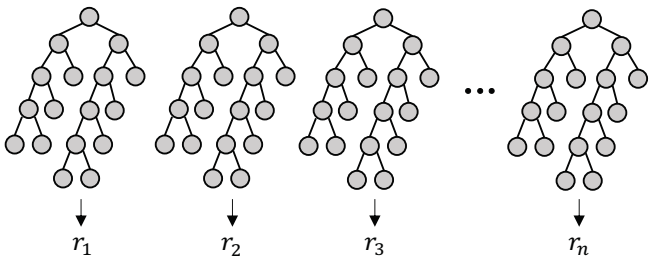
  Classification And Regression Tree (CART)
- Random Forest
- Gradient Boosting Machine and XGBoost

41

Tree-based Models

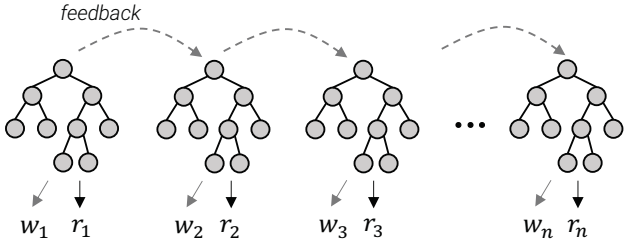
- In RF, trees are trained in
  - fully,
  - parallelly, and
  - independently of each other
- In GBM, trees are trained in
  - not fully,
  - sequentially, and
  - dependently (the next tree is trained in a way that reduces the error of previous trees)

Random Forest (RF)



Combined result: $\frac{1}{n}\sum_{i=1}^{n} r_i$

Gradient Boosting Machine (GBM)

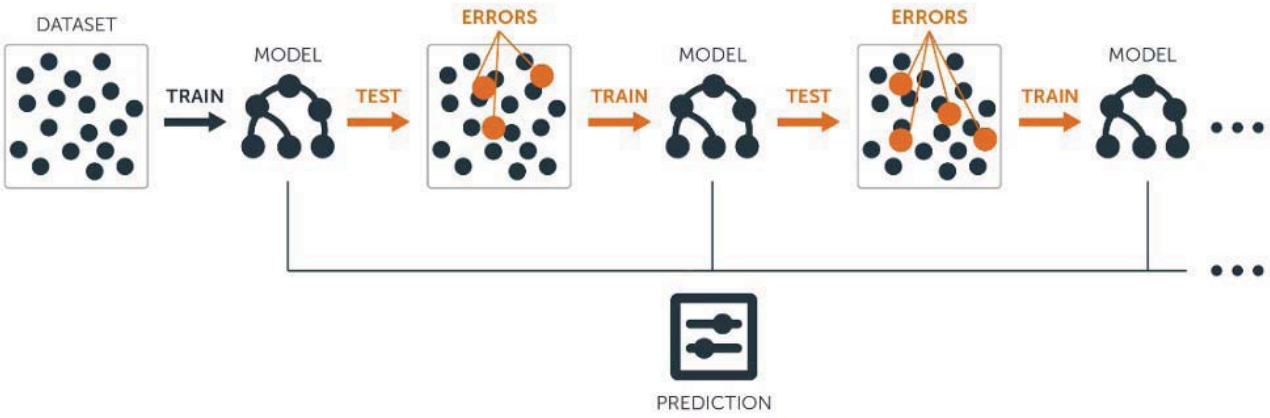


Combined result: $\sum_{i=1}^{n} w_i r_i$

42

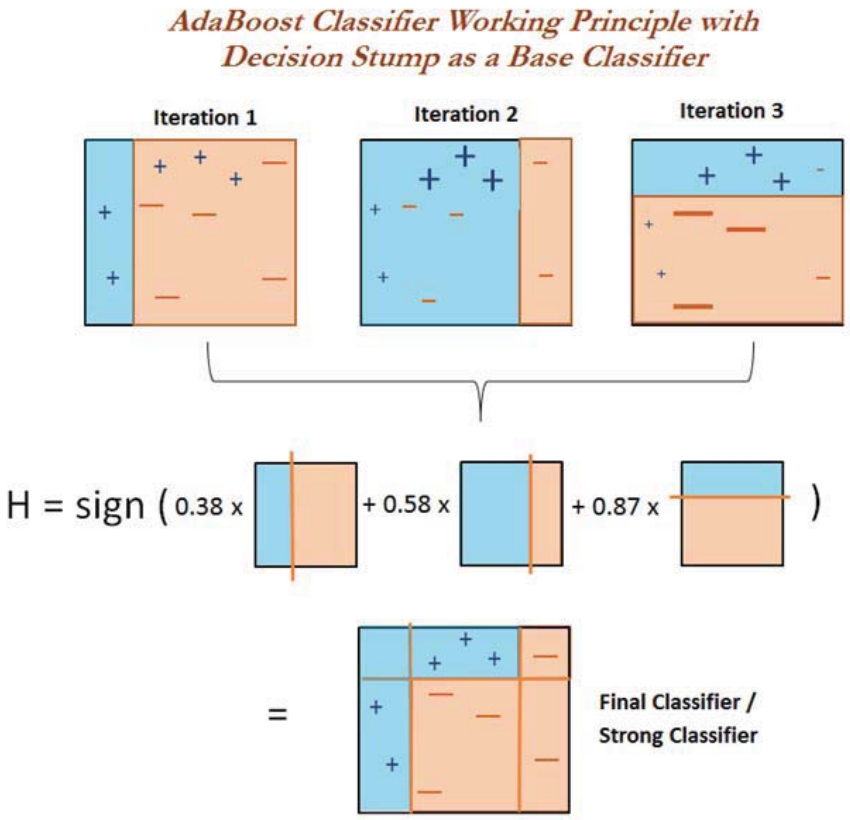
Boosting (a.k.a. additive training)

- Boosting combines weak "learners" into a single strong learner, in an iterative fashion.



43

Adaboost



44

Boosting

- Simple boosting (without weights for trees)

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i) = F_1(x_i)$$

$$\hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i) = f_1(x_i) + f_2(x_i) = F_1(x_i) + f_2(x_i) = F_2(x_i)$$

$$\dots$$

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) = \sum_{k=1}^{t-1} f_k(x_i) + f_t(x_i) = F_{t-1}(x_i) + f_t(x_i) = F_t(x_i)$$

45

Gradient boosting

- Gradient boosting method assumes a real-valued y and seeks an approximation $\hat{F}(x)$ in the form of a weighted sum of functions $h_i(x)$.
- In the training phase, we should define loss function $L(y, F(x))$.

$$\hat{F}(x) = argmin_F \mathbb{E}_{x,y}[L(y, F(x)]$$

$$F_t(x) = F_{t-1}(x) + argmin_f \sum_{i=1}^{n} L(y_i, F_{t-1}(x_i) + f(x_i))$$

$$= F_{t-1}(x) - \gamma_t \sum_{i=1}^{n} \nabla_{F_{t-1}} L(y_i, F_{t-1}(x_i))$$

$$\gamma_t = argmin_\gamma \sum_{i=1}^{n} L\left(y_i, F_{t-1}(x_i) - \gamma \frac{\partial L(y_i, F_{t-1}(x_i))}{\partial F_{t-1}(x_i)}\right)$$

46

XGBoost (eXtreme Gradient Boosting tree)

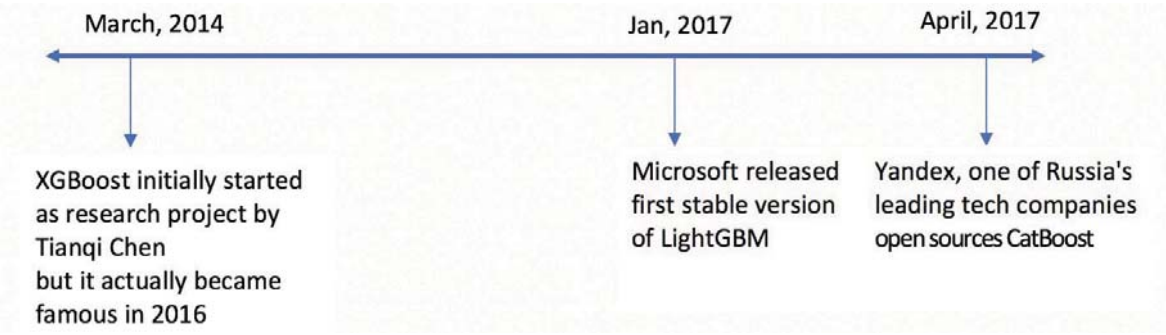
- Gradient boosting 방법의 어려움
  - 너무 많은 계산량
  - 분산 컴퓨팅이 어려움 (sequential하게 각 나무를 학습하므로)

- XGBoost의 등장 (2014 → 2016 (KDD))
  - tree를 만들어내는 과정에서 여러 근사적 알고리즘이 들어감
    - Approximate algorithm for split finding, Sparsity-aware split finding, Weighted quantile sketch, etc.
  - 연산 과정의 최적화를 통해 컴퓨팅 자원을 최대한으로 활용하는 방법 개발
    - Column block for parallel learning, Cache-aware access, Out-of-core computation, etc.

47

History of GBM

- XGBoost
- LightGBM
- CatBoost



48

Tree-based Models
- 실습

49

ICU mortality prediction

Crit Care Med. 2018 Jun;46(6):e481-e488. doi: 10.1097/CCM.0000000000003011.

Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients.

Delahanty RJ[1], Kaufman D[2], Jones SS[1].

Author information

1 Tenet Healthcare, Nashville, TN.
2 The Intensivist Group/Sound Physicians, Tacoma, WA.

Abstract

OBJECTIVES: Risk adjustment algorithms for ICU mortality are necessary for measuring and improving ICU performance. Existing risk adjustment algorithms are not widely adopted. Key barriers to adoption include licensing and implementation costs as well as labor costs associated with human-intensive data collection. Widespread adoption of electronic health records makes automated risk adjustment feasible. Using modern machine learning methods and open source tools, we developed and evaluated a retrospective risk adjustment algorithm for in-hospital mortality among ICU patients. The Risk of Inpatient Death score can be fully automated and is reliant upon data elements that are generated in the course of usual hospital processes.

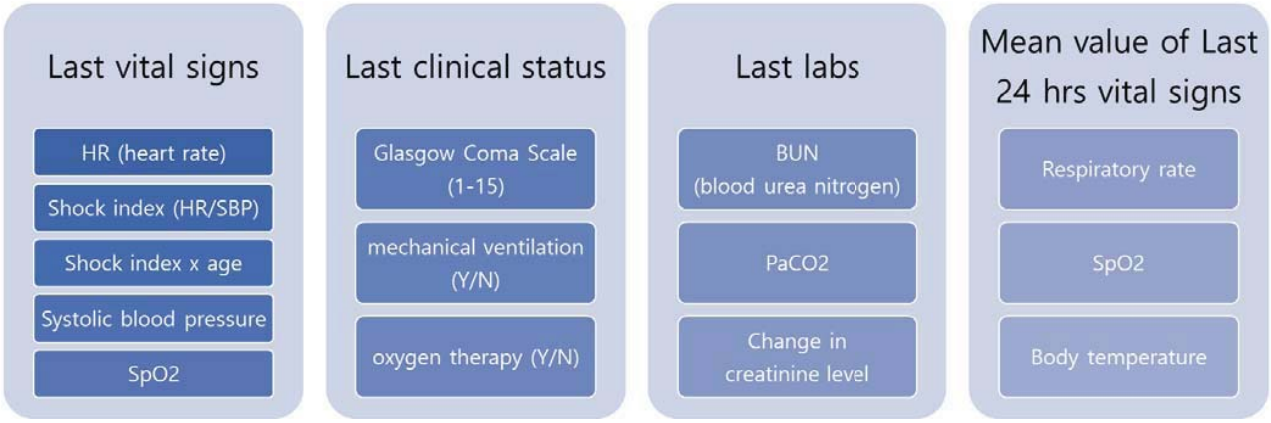
SETTING: One hundred thirty-one ICUs in 53 hospitals operated by Tenet Healthcare.

PATIENTS: A cohort of 237,173 ICU patients discharged between January 2014 and December 2016.

- ICU mortality를 예측하는 RIPD score를 제안
- 기존 방법들 (APACHE-IV, MPM-III 등) 은 AUC가 0.81인 것에 비해, machine learning 알고리즘 (XGBoost) 에 의해 계산된 RIPD score는 AUC 0.94를 기록

ICU mortality prediction

Reduced-RIPD (14 variables)

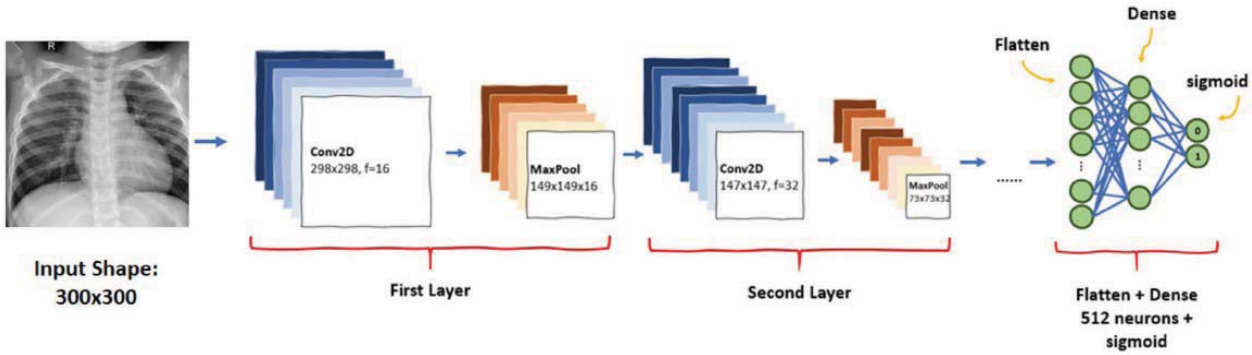


Convolutional Neural Network (CNN)

52

Convolutional Neural Network (CNN)

- 이미지, 비디오를 분석하기 위한 딥러닝 모델
- 의료 영상에 대한 인공지능 연구에서도 널리 사용되고 있음



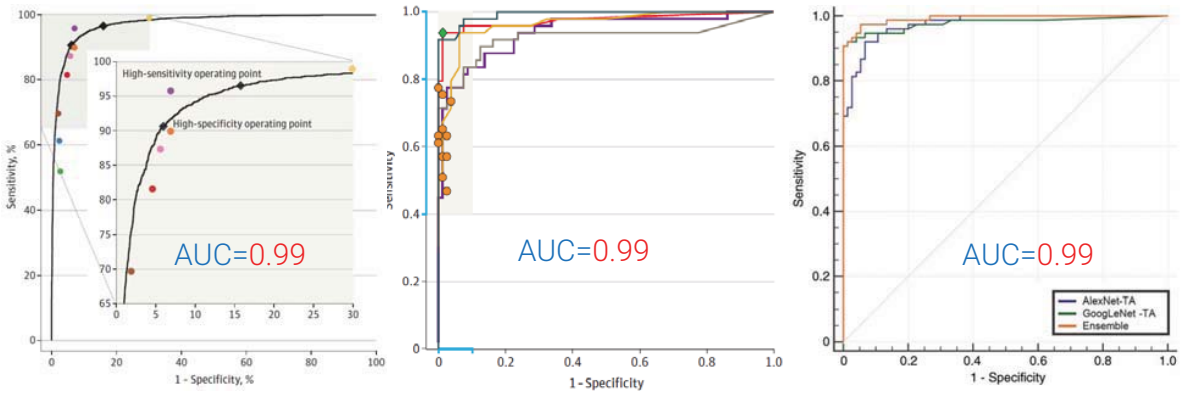
<Deep learning steps for automated detection of COVID-19>

Source: Review on COVID-19 diagnosis models based on machine learning and deep learning approaches, Alyasseri et al., Expert Systems, 2021.

53

Convolutional Neural Network (CNN)

- 2010년대 중반부터, CNN 모델이 전문의와 거의 유사한 수준의 판독 성능을 보이는 연구들이 등장



Gulshan, JAMA, 2016

[안과학] 안저사진으로 당뇨병성 망막증 판독

Bejnordi, JAMA, 2017

[병리학] 유방암의 림프절 전이 판독

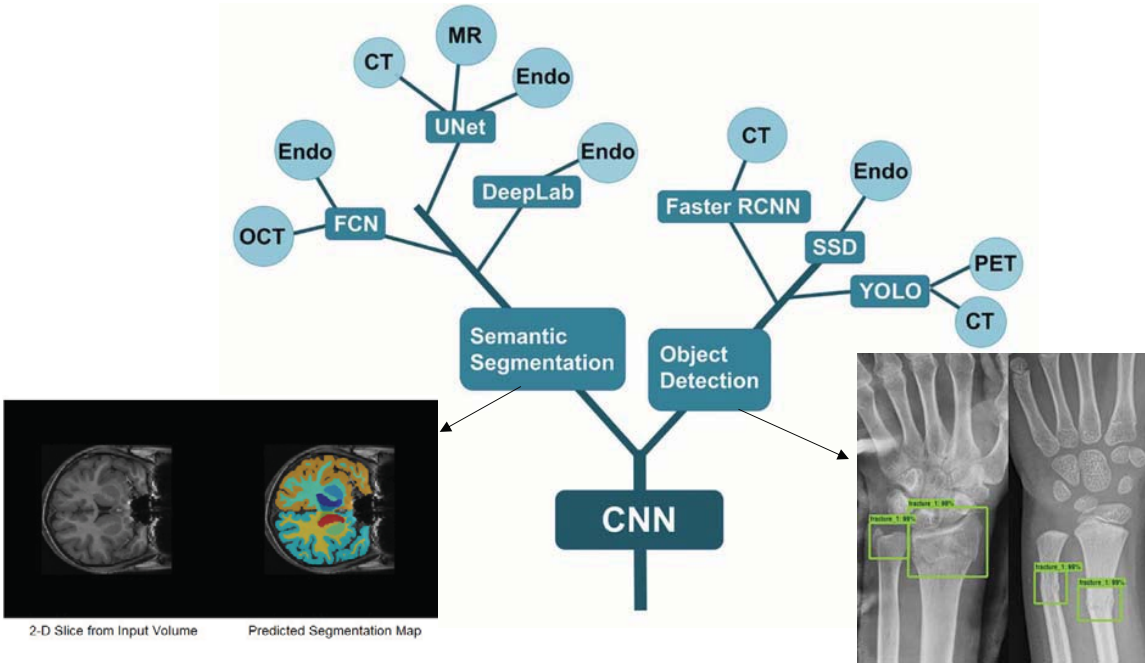
Lakhani, Radiology, 2017

[영상의학] 흉부 X-ray에서 결핵 판독

54

Convolutional Neural Network (CNN)

- 병변 탐지, 조직 분할 등에서 사용되는 여러 AI들의 뿌리



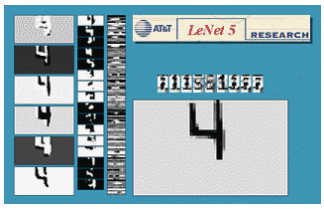
Source
- Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medica Imaging Analysis, Yang and Yu, Frontiers in Oncology, 2021
- Brain MRI Segmentation using Pretrained 3D U-Net Network (https://kr.mathworks.com/help/medical-imaging/ug/Brain-MRI-Segmentation-Using-Trained-3-D-U-Net.html)
- Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs, Thian et al., Radiology:Artificial Intelligence, 2019

55

Convolutional Neural Network (CNN)

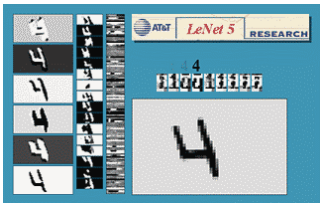
- CNN is expected to robust against the following variants.

Shift

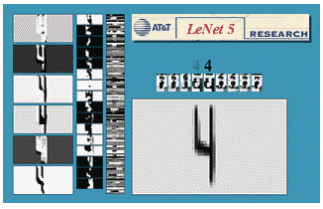


Squeezing

Scale



Stroke width



Rotation

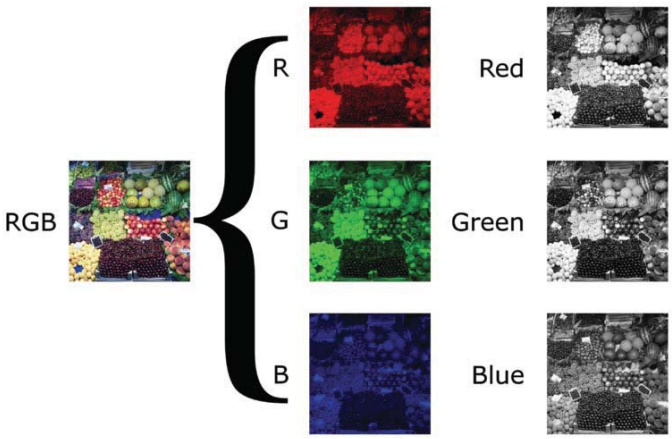
...

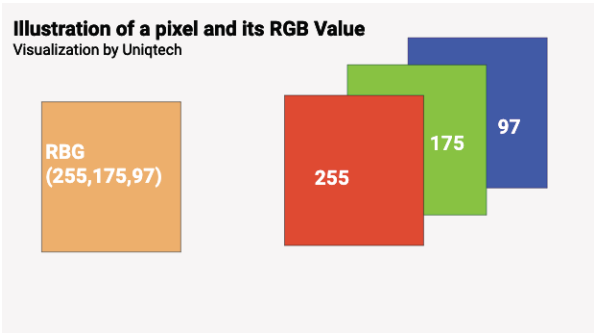
Source: Yann LeCun's homepage

56

Image Data

- Image는 3개의 채널 (Red, Green, Blue → RGB)로 구성됨



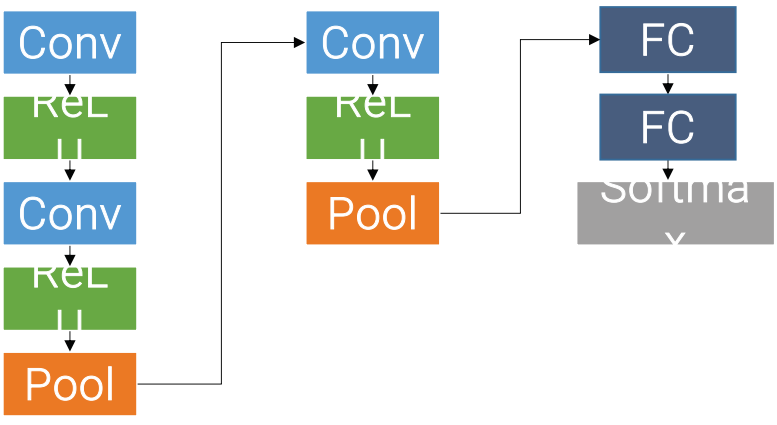


57

CNN Structures

$$[(\text{Conv} \rightarrow \text{ReLU}) * k \rightarrow \text{Pool}] * m$$

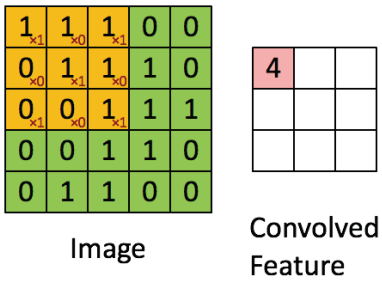
$$\rightarrow (\text{FC} \rightarrow \text{ReLU}) * n \rightarrow \text{Softmax}$$



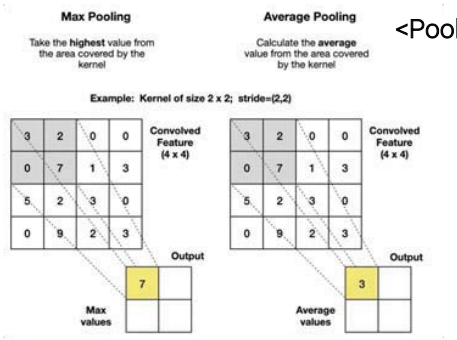
58

Convolutional Neural Network

<Convolution layer>

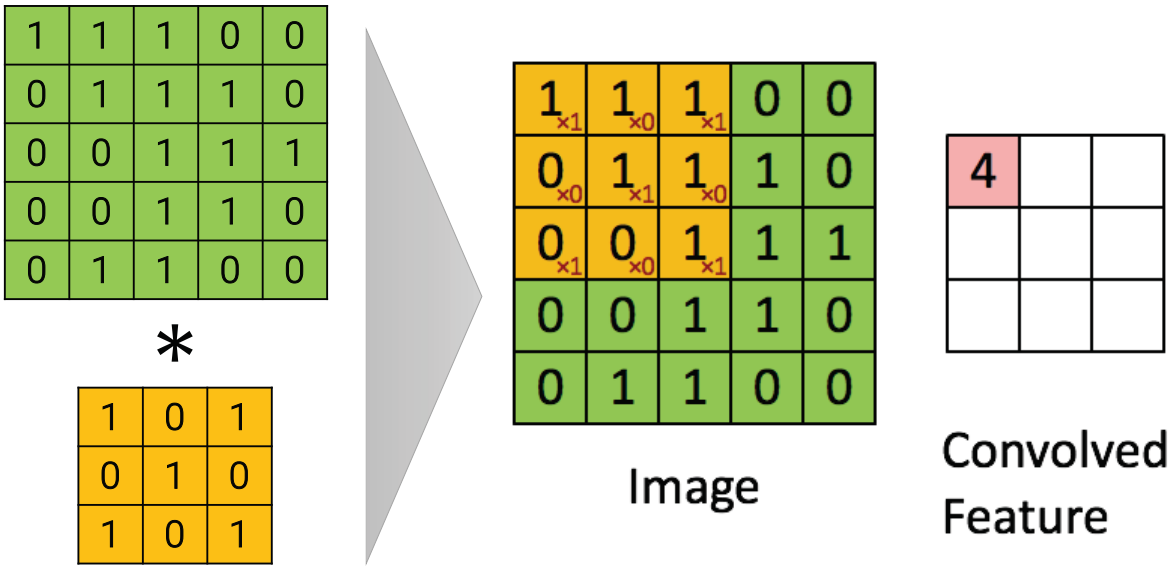


<Pooling layer>



Convolution (합성곱) for 2-D

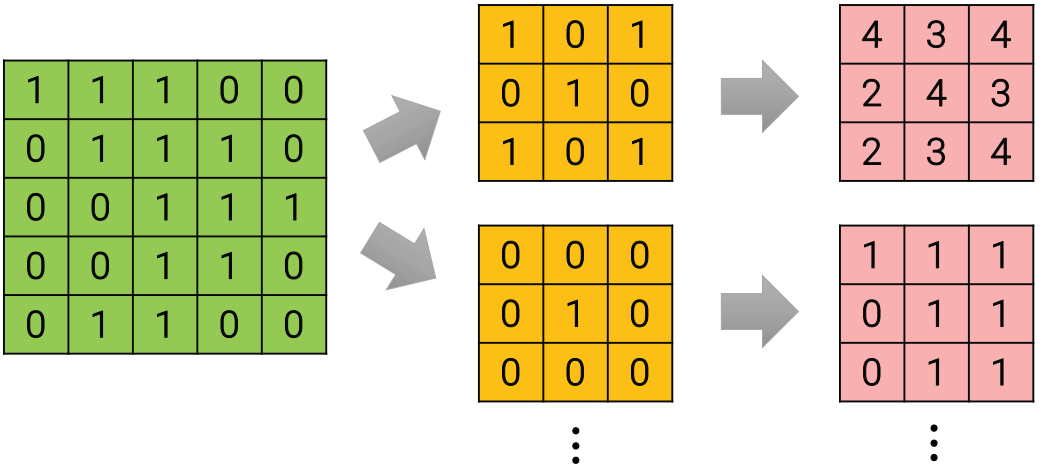
- 각 local 정보에 동일한 filter(or kernel)을 적용하여 값을 도출
  - 5x5의 텐서 데이터에 3x3의 필터를 가로, 세로 1칸씩 움직이며 합성곱을 하면,



http://deeplearning.stanford.edu/wiki/index.php/Feature_extraction_using_convolution

Convolution (합성곱) for 2-D

- 여러 개의, 각기 다른 filter를 이용
  - 5x5의 텐서 데이터에 3x3의 필터 *n*개를 가로, 세로 1칸씩 움직이며 합성곱을 하면,

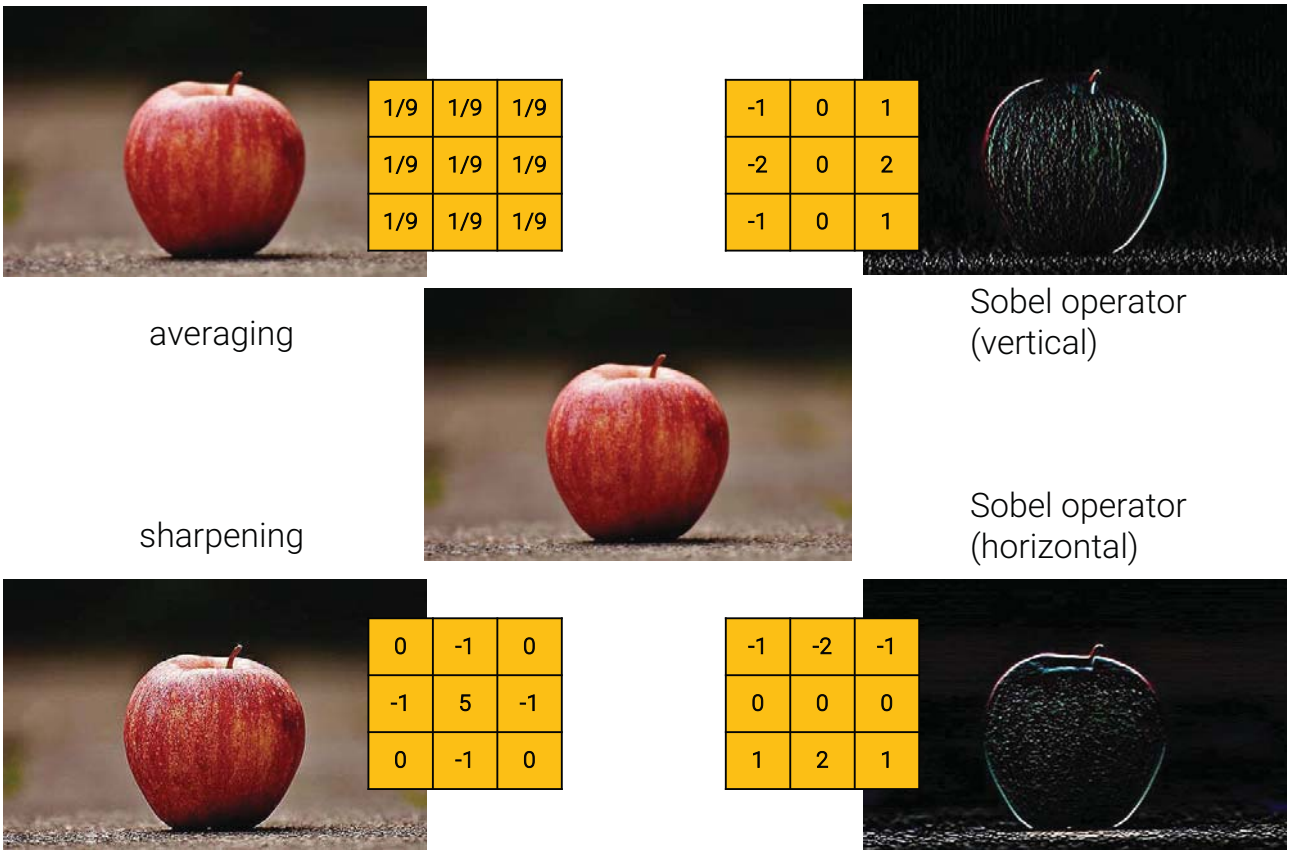


Tensor data with size 5x5x1 * *n* filters (or kernels) with size 3x3x1 = *n* convolved features with size 3x3x1

5x5x1 tensor data ➔ 3x3xn tensor data

61

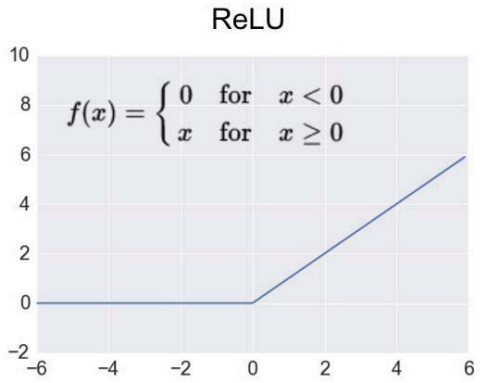
Filter의 역할 예시

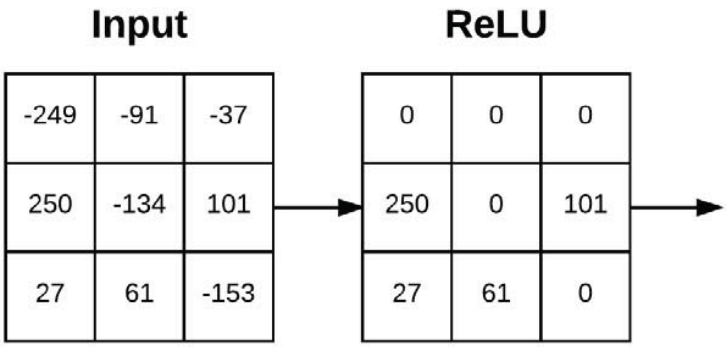


62

ReLU

- ReLU: Rectified Linear Unit
  - 인공신경망에서 사용하는 활성함수(activation function) 중 하나
  - CNN에서 많이 사용

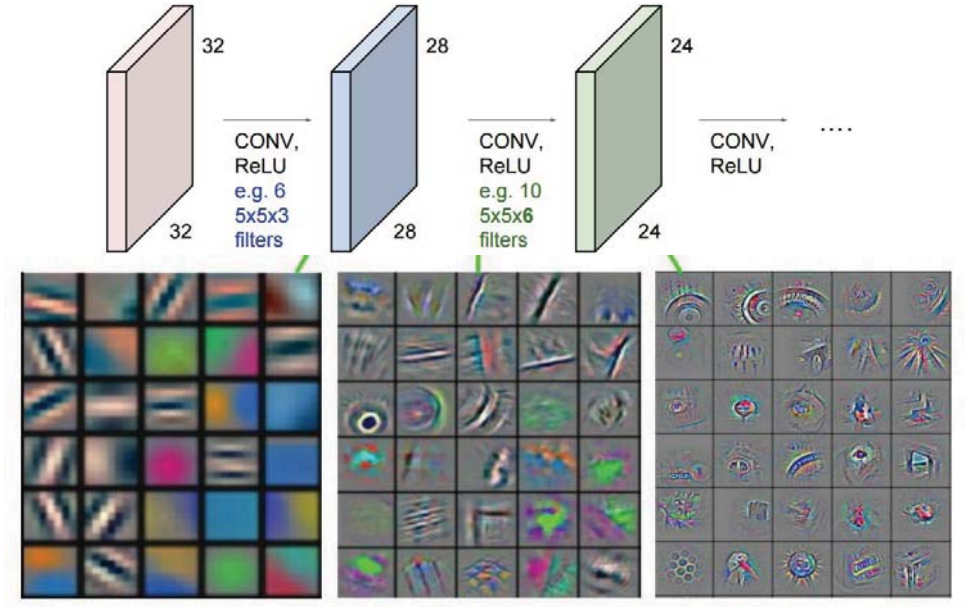




63

Conv → ReLU → Conv → ReLU → …

- 초기에는 convolutional layer 에서는 매우 간단한 feature 만 추출하지만, convolution 단계를 거듭할 수록 복잡한 feature 를 추출할 수 있다.



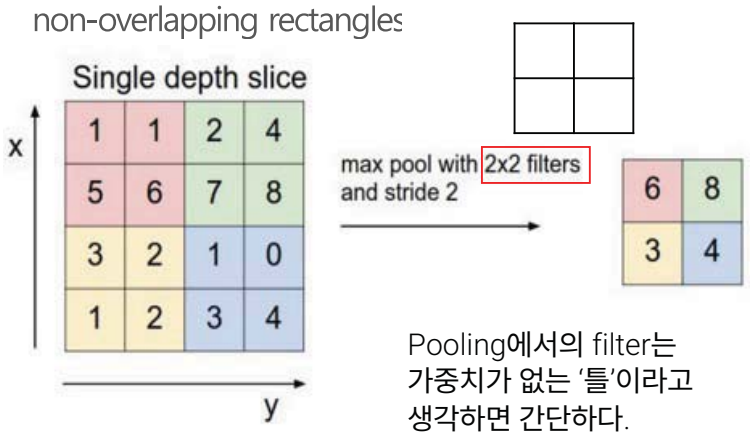
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Source : http://cs231n.stanford.edu/slides/winter1516_lecture7.pdf/

64

Pooling

- Pooling 은 유의미한 정보를 유지한 채 downsampling 기능을 함
  - Intuition : 일단 feature 를 추출했으면, 정확한 위치는 중요하지 않고, 다른 feature 에 비해 상대적 정보가 중요하다.
  - Parameter 를 줄이기 위해 representation 의 size를 줄이는 기능을 함.
  - 이를 통해 overfitting을 줄여주는 역할
  - Types of pooling
    - Average pooling
    - L2-norm pooling
    - Max pooling

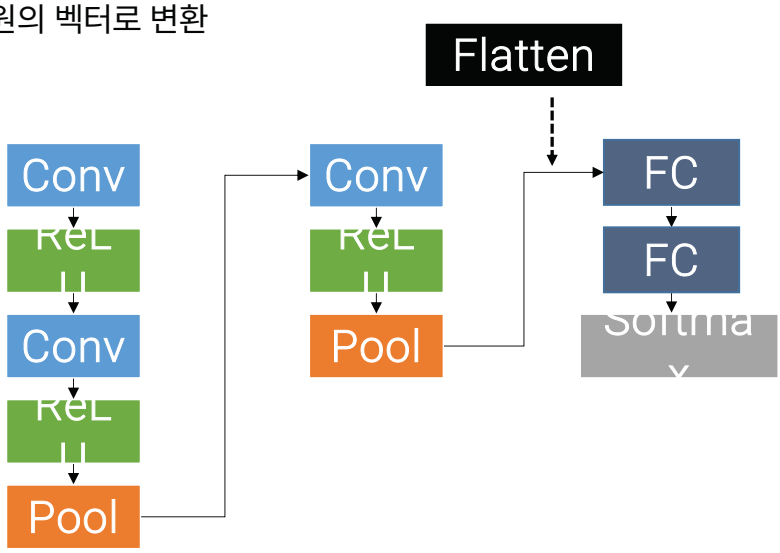


Pooling에서의 filter는 가중치가 없는 '틀'이라고 생각하면 간단하다.

65

Fully-connected

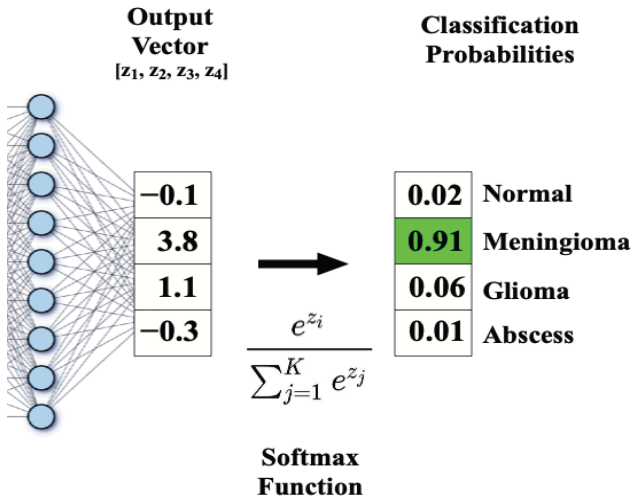
- Output layer (일반적으로 softmax) 와 convolution / pooling layer를 연결
- Activation map을 평평하게 만든 후 (flatten), 일반적은 fully-connected network와 연결
  - FC에 들어가기 전의 activation map 사이즈가 3x3x1,000 이라면, 이를 평평하게 펴서 9,000 차원의 벡터로 변환



66

Softmax

- CNN 등 여러 딥러닝 모델에서 최종 클래스의 확률을 계산하는 함수
- 전체 확률의 합을 1로 맞춰주는 역할



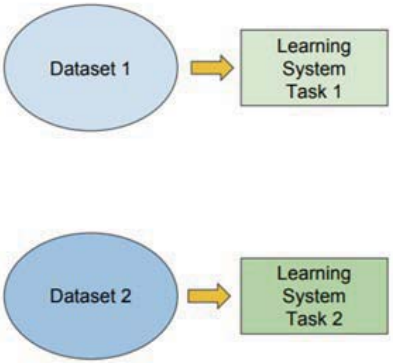
Source: https://mriquestions.com/softmax.html

Transfer Learning
(전이학습)

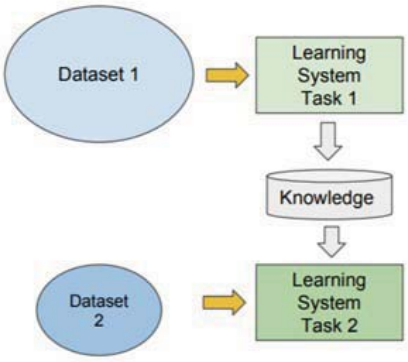
Transfer Learning (전이학습)

Traditional ML vs Transfer Learning

- Isolated, single task learning:
  - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks



- Learning of a new tasks relies on the previous learned tasks:
  - Learning process can be faster, more accurate and/or need less training data



A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning | by Dipanjan (DJ) Sarkar | Towards Data Science

69

Transfer Learning (전이학습)

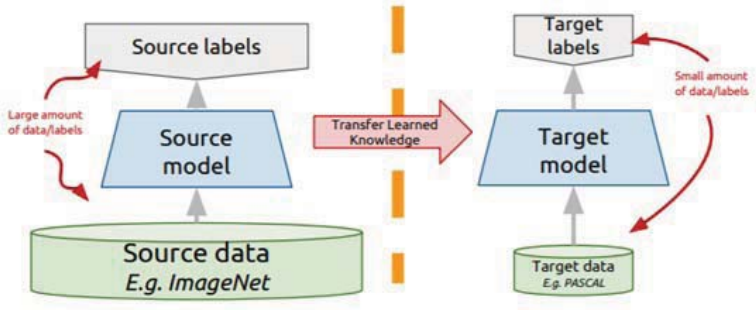
Transfer learning: idea

Instead of training a deep network from scratch for your task:

- Take a network trained on a different domain for a different **source task**
- Adapt it for your domain and your **target task**

Variations:

- Same domain, different task
- Different domain, same task



A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning | by Dipanjan (DJ) Sarkar | Towards Data Science

70

Transfer Learning의 적용 분야

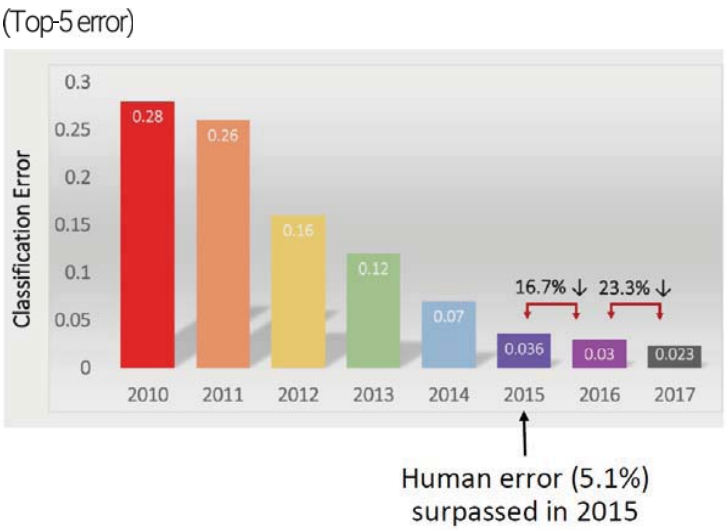
- Computer Vision
  - 대규모 데이터(image + label)를 학습한 모델(딥러닝 구조 + 학습된 weights)이 다수 존재함
  - 적용 분야: Image classification, Object detection

- Natural Language Understanding
  - 대규모 데이터(corpus)를 학습한 단어/문서 임베딩 정보들
  - BERT (Bidirectional Encoder Representations from Transformers) 를 시작으로 다양한 Language Model들이 transfer learning이 적극 활용되고 있음

71

There are so many great pretrained CNNs.

- ImageNet Challenge



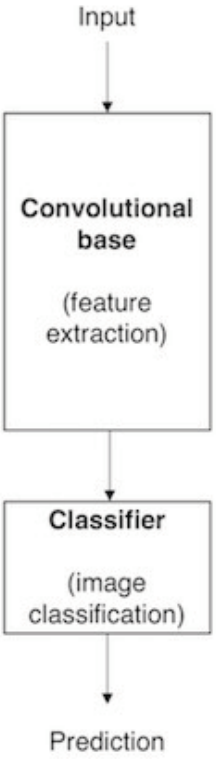
- **AlexNet (2012): First CNN (15.4%)**
  - 8 layers
  - 61 million parameters
- **ZFNet (2013): 15.4% to 11.2%**
  - 8 layers
  - More filters. Denser stride.
- **VGGNet (2014): 11.2% to 7.3%**
  - Beautifully uniform:
    3x3 conv, stride 1, pad 1, 2x2 max pool
  - 16 layers
  - 138 million parameters
- **GoogLeNet (2014): 11.2% to 6.7%**
  - Inception modules
  - 22 layers
  - 5 million parameters
    (throw away fully connected layers)
- **ResNet (2015): 6.7% to 3.57%**
  - More layers = better performance
  - 152 layers
- **CUImage (2016): 3.57% to 2.99%**
  - Ensemble of 6 models
- **SENet (2017): 2.99% to 2.251%**
  - Squeeze and excitation block: network is allowed to adaptively adjust the weighting of each feature map in the convolutional block.

Source: Deep Learning: State of the Art, Lex Fridman, MIT, 2019.

72

Transfer learning in CNN

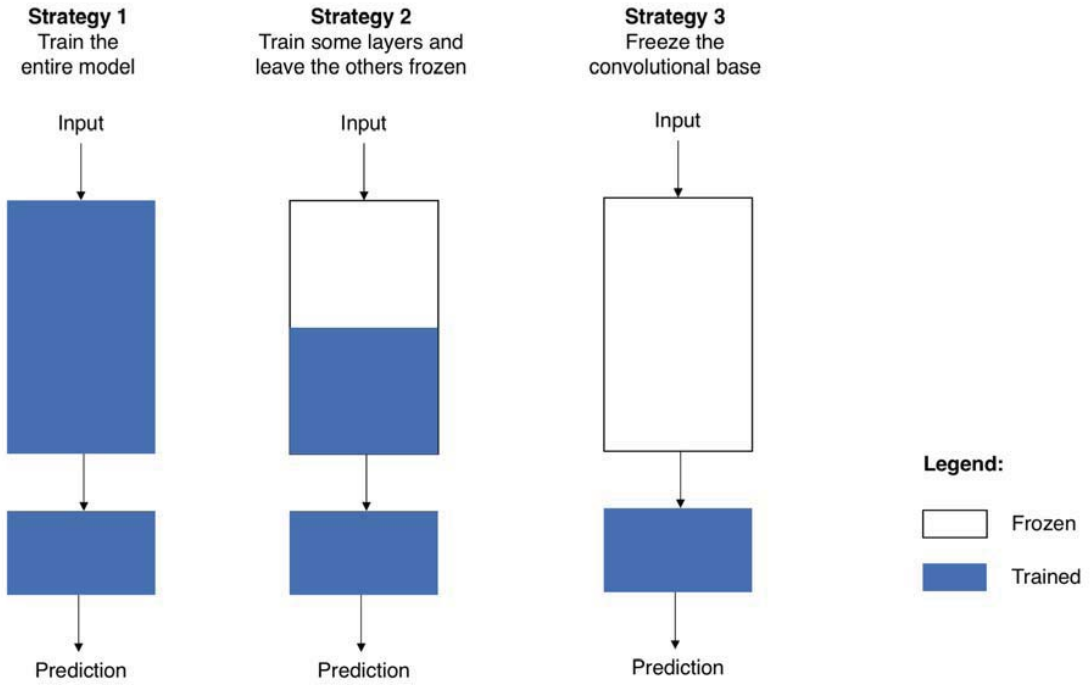
- Convolutional base
  - Structure: Stack of convolution and pooling layers
  - Goal: Generate features from the image
- Classifier
  - Structure: fully-connected network *or* other ML algorithms such as SVM, Random Forest, Gradient Boosting Machine, etc.
  - Goal: Classify the image using features



Transfer learning from pre-trained models | by Pedro Marcelino | Towards Data Science

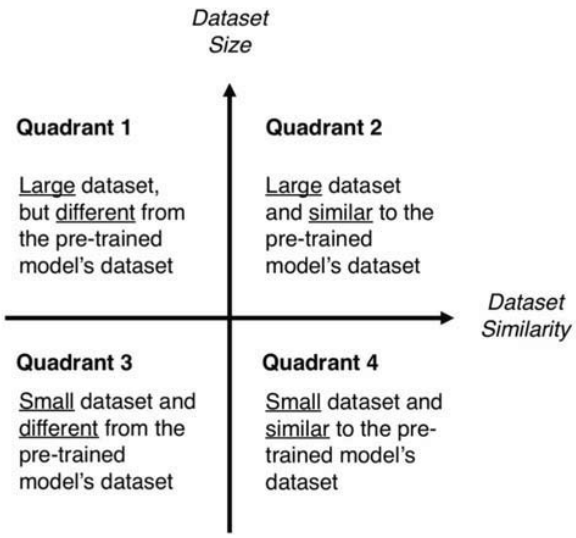
Transfer learning in CNN

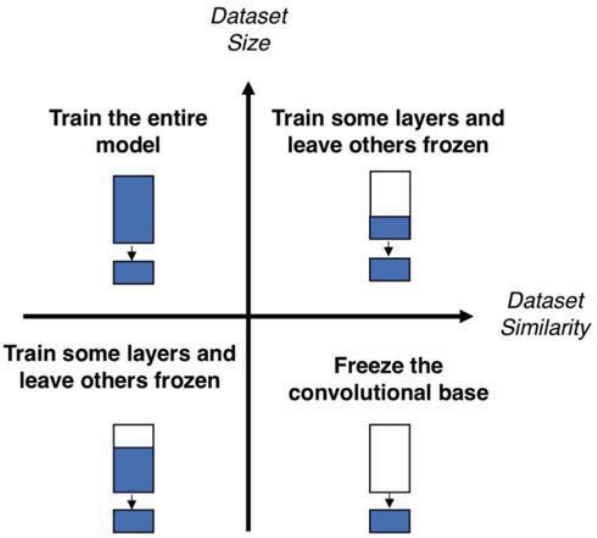
- How to fine-tune the pretrained model: 3 strategies



Transfer learning from pre-trained models | by Pedro Marcelino | Towards Data Science

Transfer learning in CNN



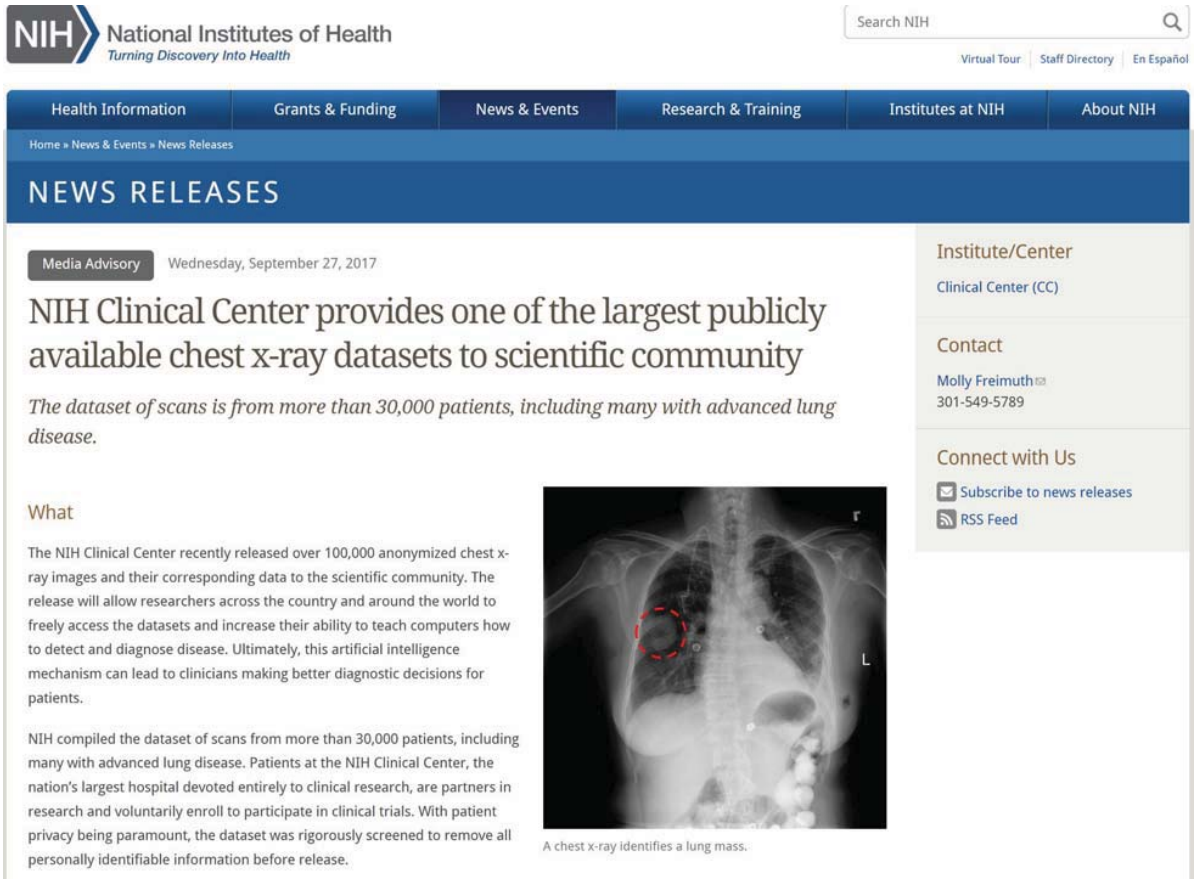


Transfer learning from pre-trained models | by Pedro Marcelino | Towards Data Science

75

실습: NIH Chest X-ray Dataset (ChestX-ray14)

76



NIH National Institutes of Health
Turning Discovery Into Health

Search NIH

Virtual Tour | Staff Directory | En Español

Health Information | Grants & Funding | News & Events | Research & Training | Institutes at NIH | About NIH

Home » News & Events » News Releases

NEWS RELEASES

Media Advisory Wednesday, September 27, 2017

NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community

The dataset of scans is from more than 30,000 patients, including many with advanced lung disease.

Institute/Center

Clinical Center (CC)

Contact

Molly Freimuth
301-549-5789

Connect with Us

Subscribe to news releases

RSS Feed

What

The NIH Clinical Center recently released over 100,000 anonymized chest x-ray images and their corresponding data to the scientific community. The release will allow researchers across the country and around the world to freely access the datasets and increase their ability to teach computers how to detect and diagnose disease. Ultimately, this artificial intelligence mechanism can lead to clinicians making better diagnostic decisions for patients.

NIH compiled the dataset of scans from more than 30,000 patients, including many with advanced lung disease. Patients at the NIH Clinical Center, the nation's largest hospital devoted entirely to clinical research, are partners in research and voluntarily enroll to participate in clinical trials. With patient privacy being paramount, the dataset was rigorously screened to remove all personally identifiable information before release.

A chest x-ray identifies a lung mass.

https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

77

NIH Chest X-ray Dataset

- Full dataset
  - https://www.kaggle.com/datasets/nih-chest-xrays/data
  - 112,120 x-ray images with disease labels from 30,805 unique patients
- Random sample dataset
  - https://www.kaggle.com/datasets/nih-chest-xrays/sample
  - 5% random sample: 5,606 images

78

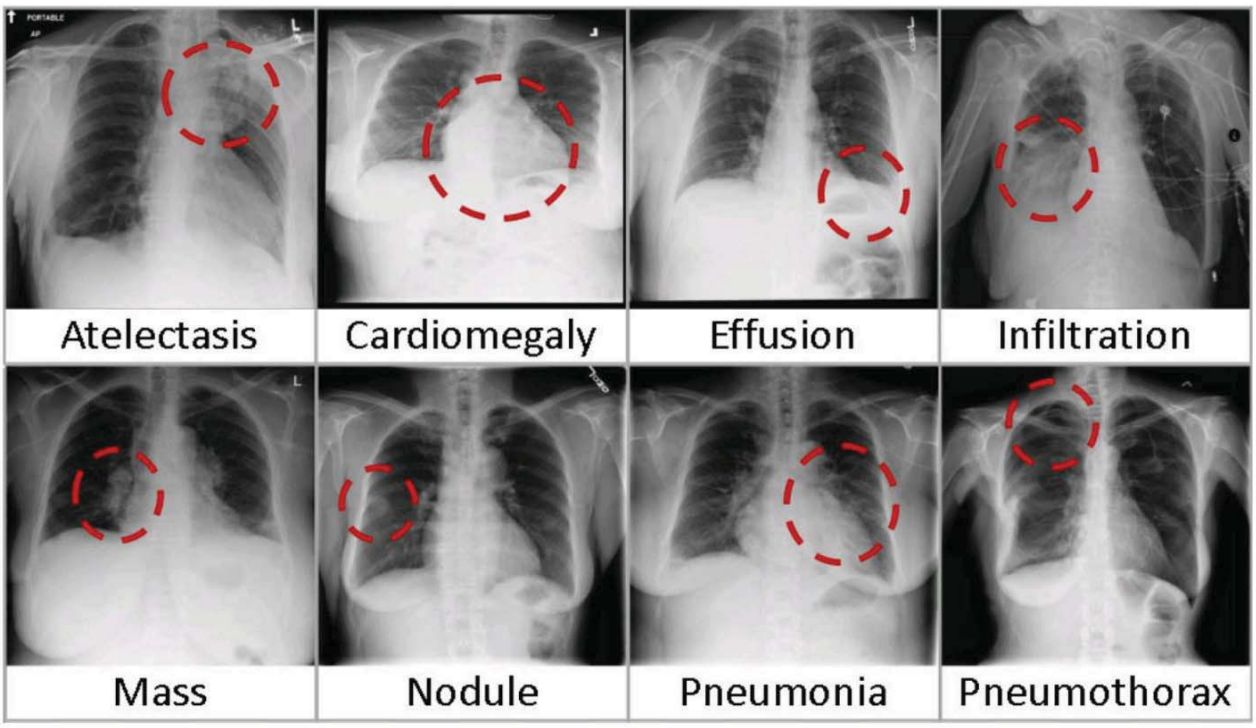
NIH Chest X-ray Dataset

- There are 15 classes (14 diseases, and one for "No findings").
  Images can be classified as "No findings" or one or more disease classes:
  - Atelectasis
  - Consolidation
  - Infiltration
  - Pneumothorax
  - Edema
  - Emphysema
  - Fibrosis
  - Effusion
  - Pneumonia
  - Pleural_thickening
  - Cardiomegaly
  - Nodule Mass
  - Hernia

79

NIH Chest X-ray Dataset

B. Eight visual examples of common thorax diseases



80

NIH Chest X-ray Dataset

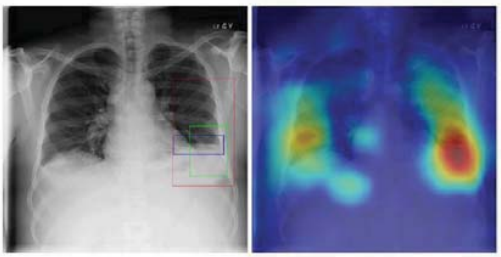
Radiology report	Keyword	Localization Result
findings include: 1. left basilar atelectasis/consolidation. 2. prominent hilum (mediastinal adenopathy). 3. left pic catheter (tip in atriocaval junction). 4. stable, normal appearing cardiomediastinal silhouette. impression: small right pleural effusion otherwise stable abnormal study including left basilar infiltrate/atelectasis, prominent hilum, and position of left pic catheter (tip atriocaval junction).	Effusion; Infiltration; Atelectasis	

Table 4. A sample of chest x-ray radiology report, mined disease keywords and localization result from the "Atelectasis" Class. Correct bounding box (in green), false positives (in red) and the ground truth (in blue) are plotted over the original image.

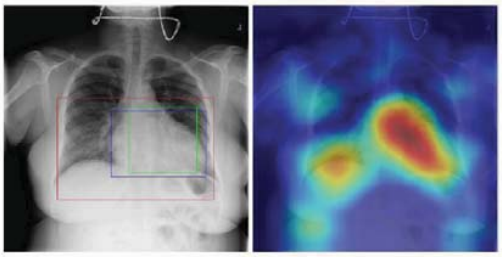
Radiology report	Keyword	Localization Result
findings include: 1. cardiomegaly (ct ratio of 17/30). 2. otherwise normal lungs and mediastinal contours. 3. no evidence of focal bone lesion. dictating	Cardiomegaly	

Table 5. A sample of chest x-ray radiology report, mined disease keywords and localization result from the "Cardiomegaly" Class. Correct bounding box (in green), false positives (in red) and the ground truth (in blue) are plotted over the original image.

81

NIH Chest X-ray Dataset

- Limitations:
  - 1) The image labels are NLP extracted so there would be some erroneous labels, but the NLP labelling accuracy is estimated to be >90%.
  - 2) Very limited numbers of disease region bounding boxes.
  - 3) Chest x-ray radiology reports are not anticipated to be publicly shared. Parties who use this public dataset are encouraged to share their "updated" image labels and/or new bounding boxes in their own studied later, maybe through manual annotation.

82

Recurrent Neural Network

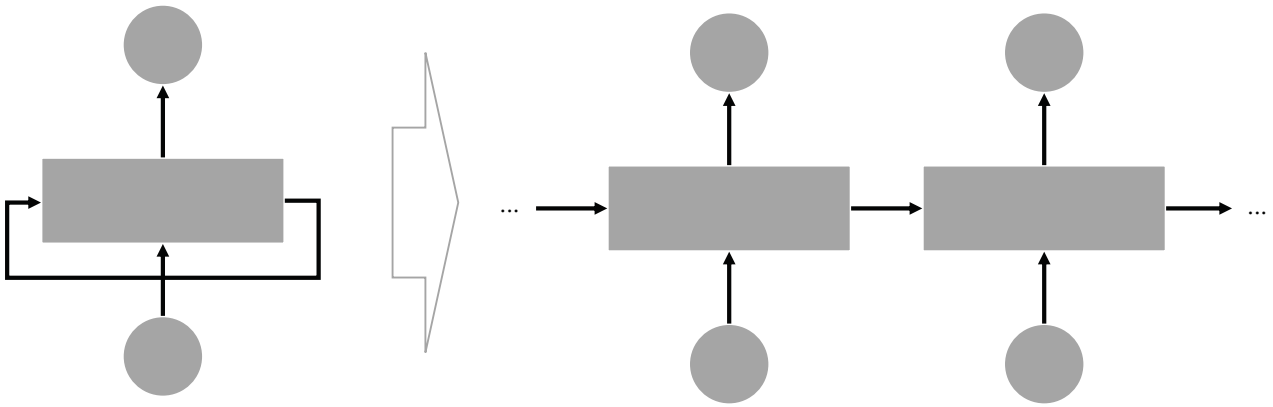
Recurrent Neural Network

- 순서가 있는 데이터 (sequential data) 를 학습하기 위한 딥러닝 모델
- 순서가 있는 데이터 예시
  - 자연어: After warming the pizza in the oven, I ate it.
  - 비디오
  - 환자의 follow-up 데이터
  - CT, MRI: axial, coronal, sagittal

$$(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)} \ldots, \boldsymbol{x}^{(T)})$$

where each data point $\boldsymbol{x}^{(t)}$, which is observed at *time step* $\boldsymbol{t}$, is a real value or real-valued vector/tensor.

Recurrent Neural Network



85

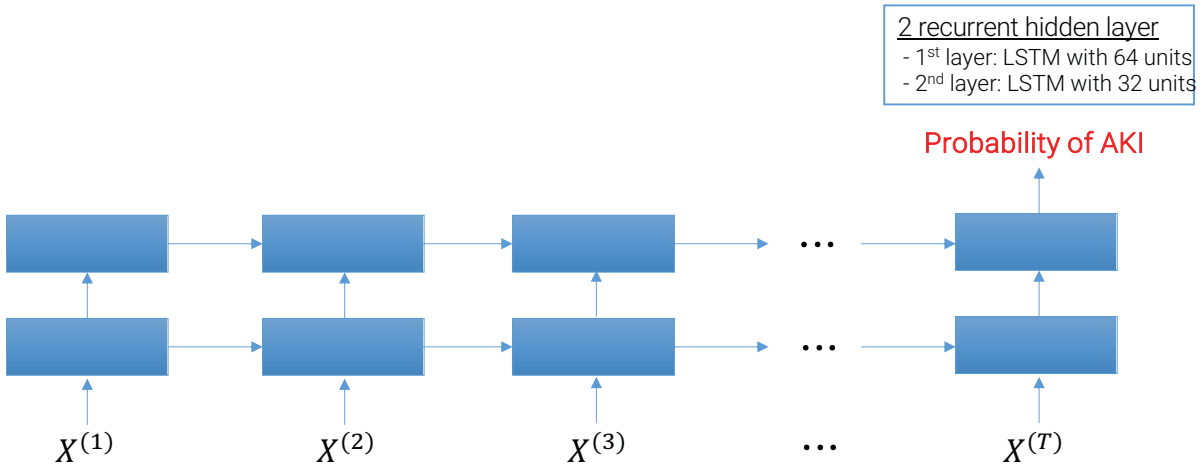
Predicting AKI using intra-OP data

- 급성 신손상 (Acute Kidney Injury)
  - 급성 신손상(acute kidney injury, AKI)은 신기능의 급격한 저하를 특징으로 하는 질환으로 환자의 이환율과 사망률을 증가시킬 뿐만 아니라, 이에 드는 비용 또한 매우 큰 편으로 임상적으로 매우 중요한 문제 (김범석, 2012)
  - AKI는 중환자실 치료를 받는 환자에서 가장 흔하게 발생하는 질환 중 하나로 미국의 경우 매년 약 20만명의 신환이 발생하는 것으로 알려져 있음
- Motivation
  - 수술 중 환자로부터 계측되는 vital signs만으로 AKI의 발생을 선제적으로 예측할 수 있는가?
  - 실시간 scoring이 가능한가?

86

Recurrent Neural Network (RNN)

- t번째 환자의 수술기록을 $X^{(t)}$ $(t = 1, 2, \ldots, T)$
- $X^{(T)}$가 투입되는 시점에 해당 환자의 AKI 발병 확률



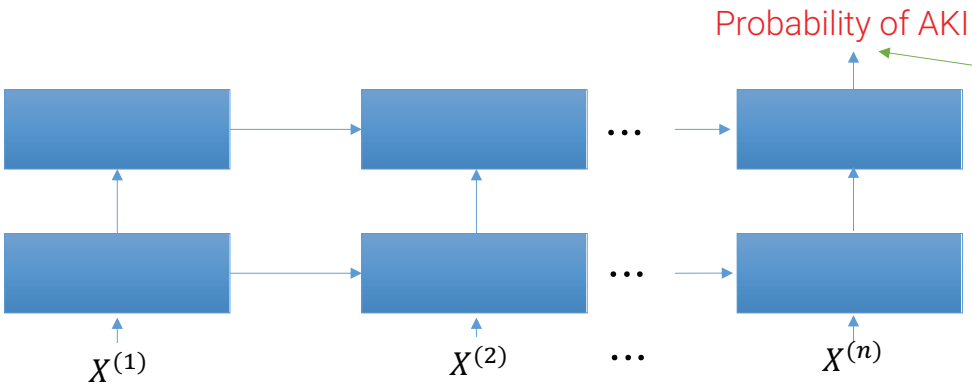
SBP(수축기혈압), MBP(평균혈압), DBP(이완기혈압),
SpO2 (혈중산소포화도), CVP(중심정맥압) 등

데이터 처리

- 다음 사항 중 적어도 하나 해당하는 레코드는 삭제
  - 측정항목 별로 값이 지나치게 크거나 작은 값을 가진 경우를 제외
  - E.g., SBP 300 이상, DBP 200 이상, MBP 267 이상, SpO2 100 이상
- RNN을 학습하기 위한 Training set과
  학습된 모델의 성능을 검증하기 위한 Test set으로 데이터 분할
  - Training set의 수술 수: 40,882 (AKI 2,190, No AKI 38,692)
  - Test set의 수술 수: 10,221 (AKI 548, No AKI 9,673)
- 데이터 정규화: 모든 값을 0과 1사이로 정규화

88

모델 고도화 진행 중



- 연령, 성별, 당뇨, 실제 수술 시간
- 수술 전 SBP, DBP, EFGR, HB, ALB, NA, K, UALB_C, RASB 약제 사용 여부
- 기저 신손상이 있었던 경우
- ...

$X^{(t)}$: 수술 중 t번째 측정된 데이터
- SBP, SBP, MBP, SpO2, CVP
- 각 측정항목이 측정된 여부
- $X^{(t-1)}$과 $X^{(t)}$ 사이에 걸린 시간

파란색 층: 2 recurrent hidden layers
- 1st layer: LSTM with 64 units
- 2nd layer: LSTM with 32 units

89

실험 결과

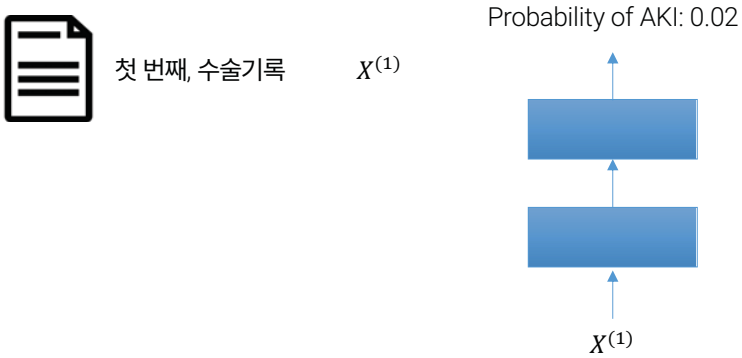
- 분류 정확도 (Accuracy): 93.98%
- 민감도 (Sensitivity): 0.6058
- 정밀도 (Precision): 0.4542

- 환자에 대한 정보와 pre-operative data를 함께 학습하는 것은 매우 효과적임
- Inter-operative data에 대한 더 세밀한 처리, 데이터 항목 추가를 통해 모델 성능 향상 도모
- Class imbalance를 해결하는 방안?

90

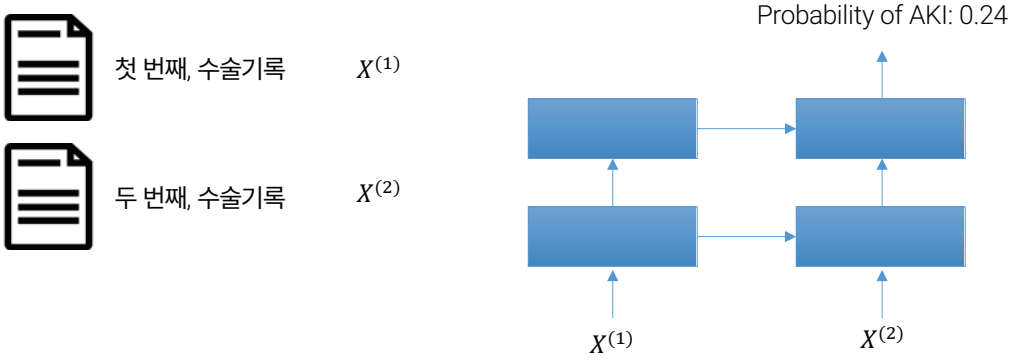
RNN이 학습이 잘 될 경우,

- 향후 수술환자의 AKI 가능성을 실시간으로 계산 가능



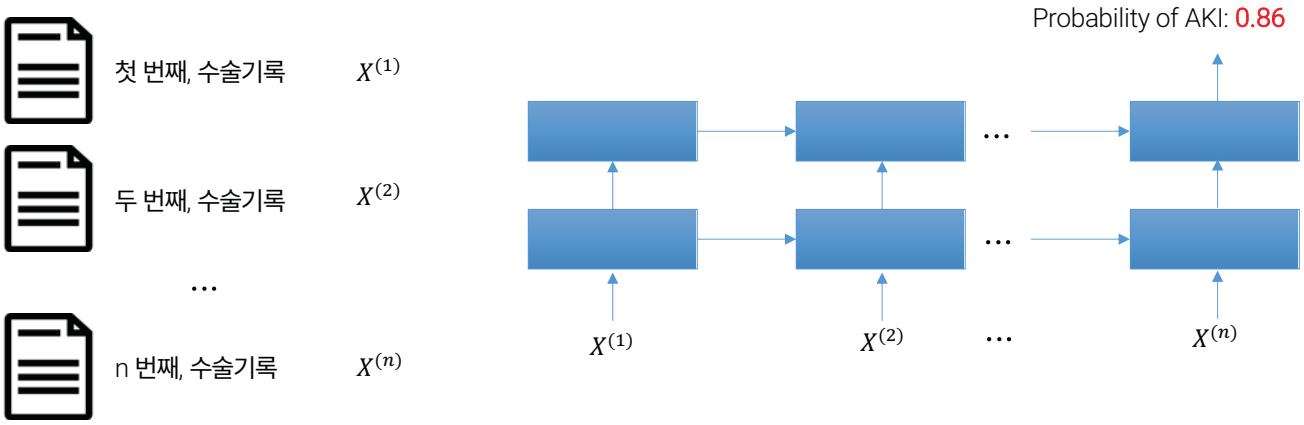
RNN이 학습이 잘 될 경우,

- 향후 수술환자의 AKI 가능성을 실시간으로 계산 가능



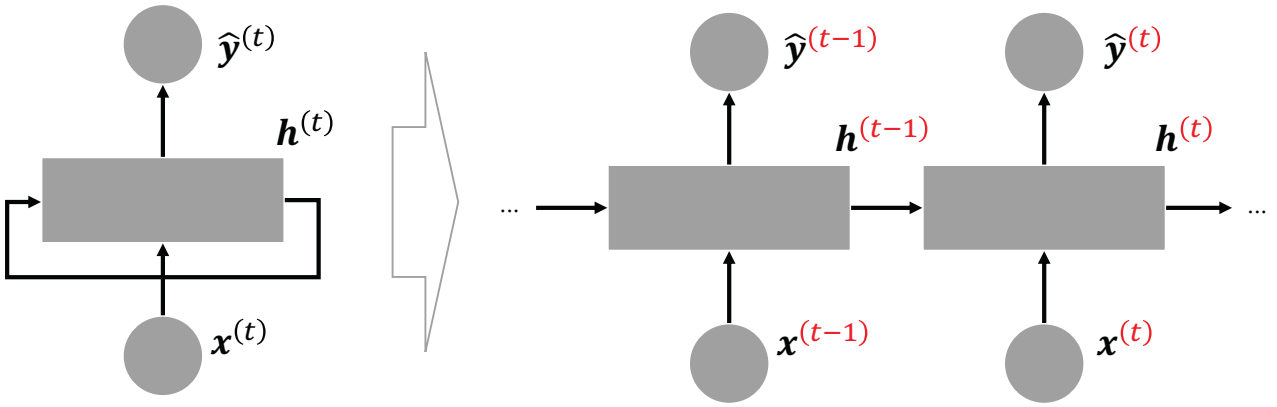
RNN이 학습이 잘 될 경우,

- 향후 수술환자의 AKI 가능성을 실시간으로 계산 가능



Recurrent neural network (RNN)

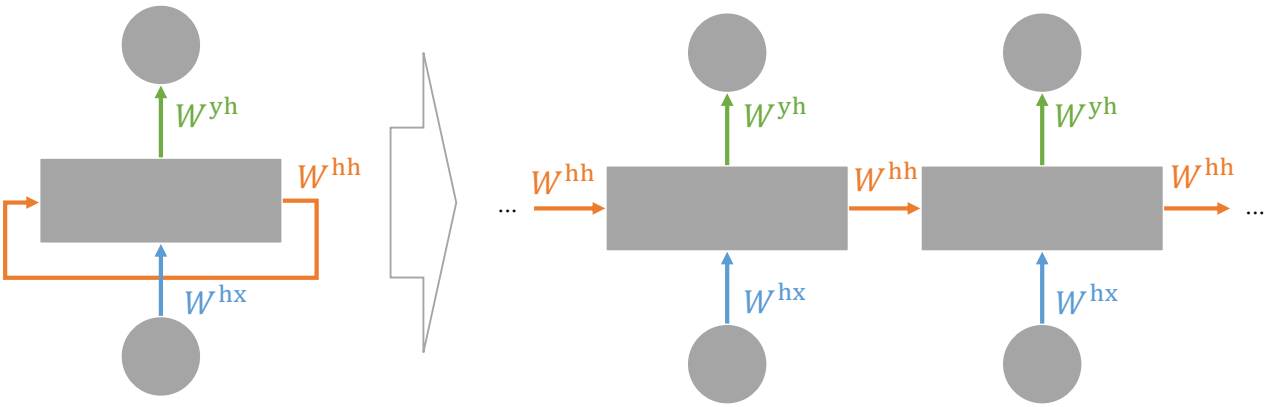
- The data points, hidden node values and outputs: Time-variant



94

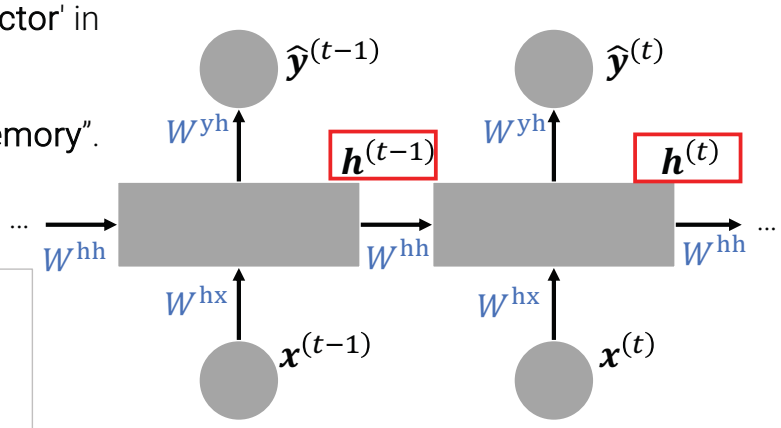
Recurrent neural network (RNN)

- Weights and biases: Time-invariant.



Recurrent neural network (RNN)

- State (vector): $\boldsymbol{h}^{(t)}$
  - RNNs maintain a '**state vector**' in their each hidden layer.
  - It is considered as the "**memory**".



<Recurrent neural network>

The states (are expected to) contain information about the history of all the past elements of the sequence.

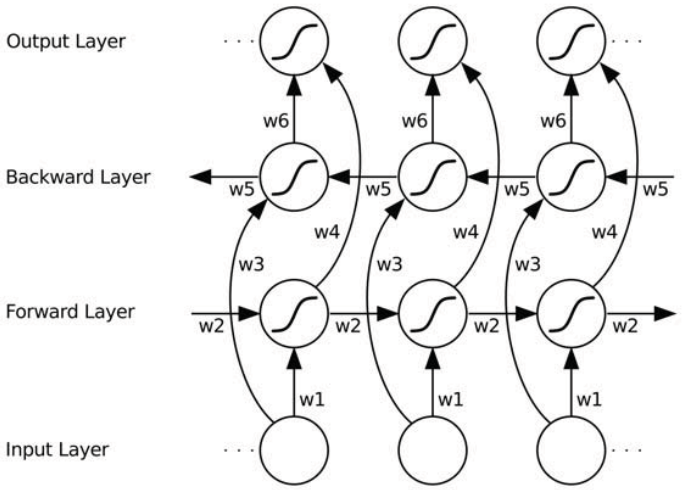
RNNs are better for tasks that involve sequential inputs such as speech, language, and genome.

$$\boldsymbol{h}^{(t)} = f(W^{hx}\boldsymbol{x}^{(t)} + W^{hh}\boldsymbol{h}^{(t-1)} + b^{h})$$

$$\hat{\boldsymbol{y}}^{(t)} = g(W^{yh}\boldsymbol{h}^{(t)} + b^{y})$$

Bidirectional RNN

- 본래 순서의 시퀀스와 역순서의 시퀀스에 대한 각각의 RNN을 만든 후, 두 RNN의 출력층을 통합
- 현재 시점의 출력값에 대해 과거와 미래의 state가 모두 반영됨
  - 더 많은 정보가 이용된다는 측면에서, 본래 순서의 시퀀스만 사용한 RNN보다 더 좋은 효과가 있을 수 있음

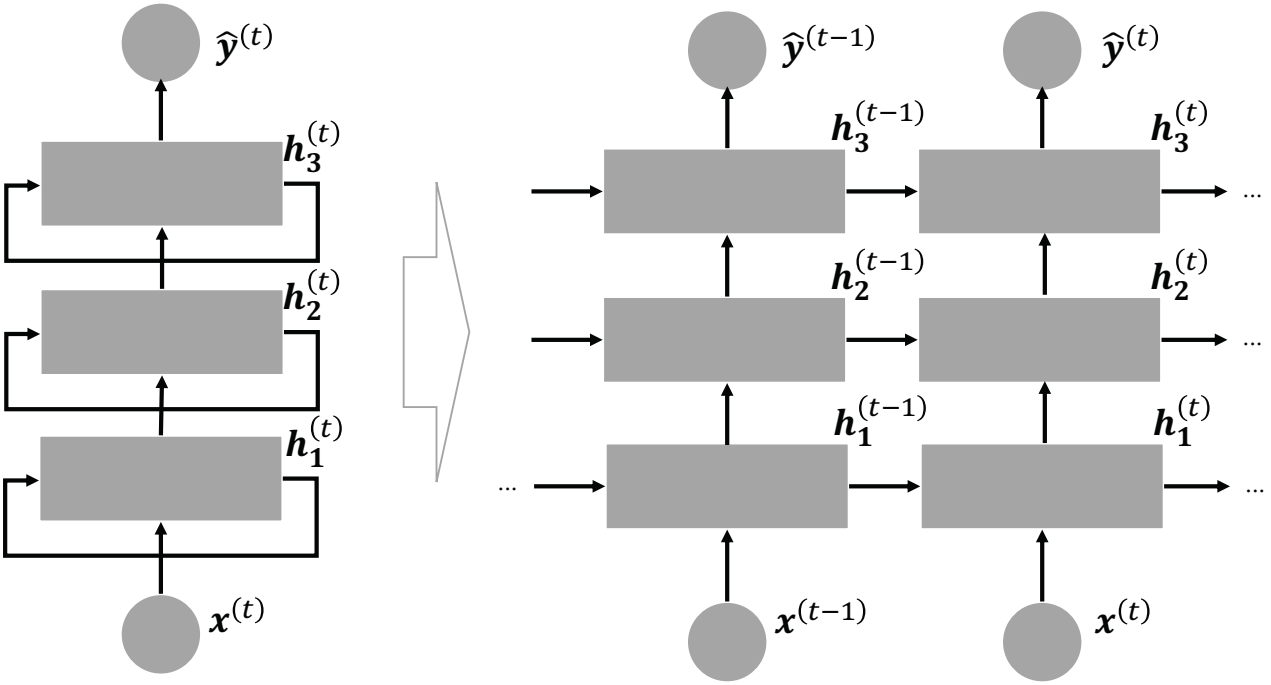


Alex Graves, "Supervised sequence labeling with recurrent neural network"

97

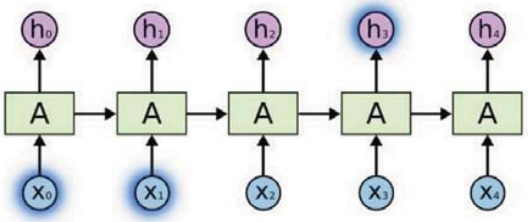
Deep recurrent neural net

- 순환하는 은닉층을 다수 쌓음으로써 신경망구조를 더욱 깊게 만들 수 있음



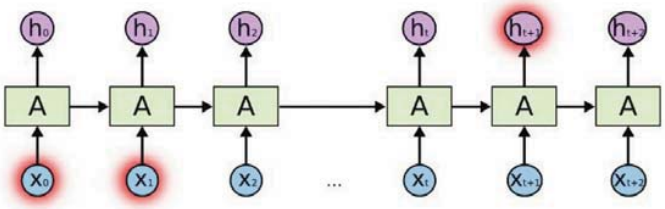
98

Long-Term Dependency



- Short-term dependence:
  Bob is eating an **apple.**
- Long-term dependence:

Context → **Bob** likes **apples**. He is hungry and decided to have a snack. So now he is eating an **apple.**

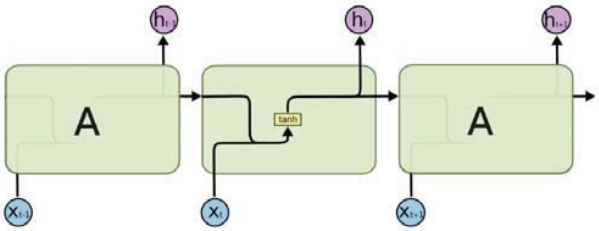


In theory, vanilla RNNs can handle arbitrarily long-term dependence.

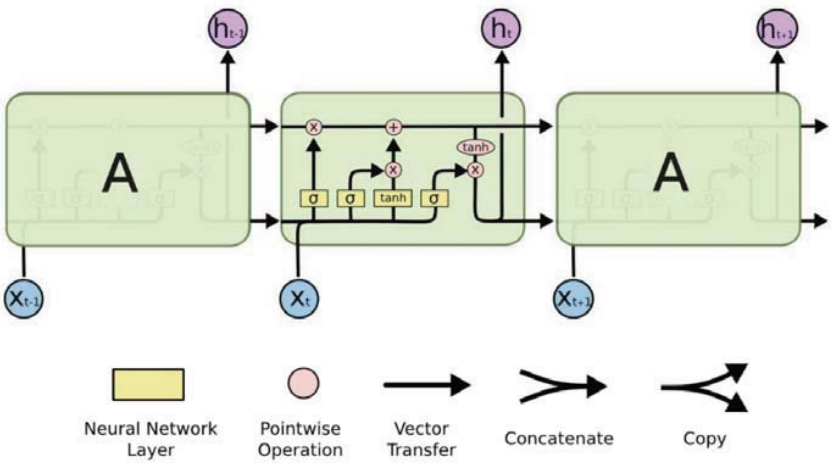
In practice, it's difficult.

Long Short Term Memory (LSTM) Networks

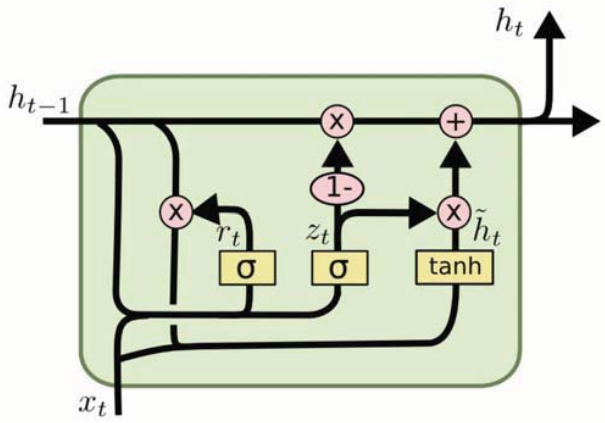
Vanilla RNN:



LSTM:



Gated recurrent units (GRUs)



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

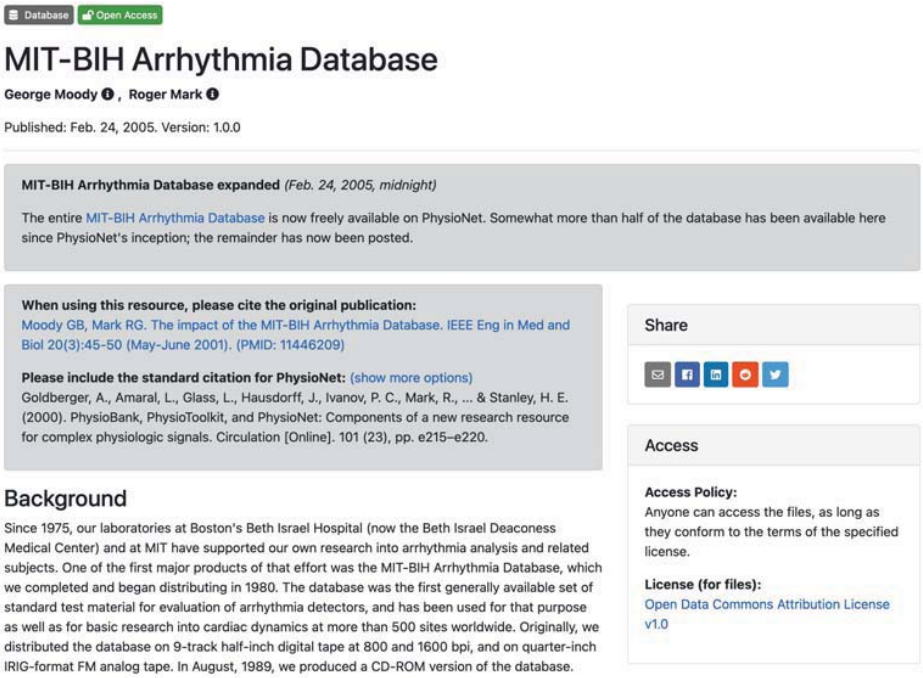
101

실습: MIT-BIH Arrhythmia Database

102

MIT-BIH Arrhythmia Database

- https://physionet.org/content/mitdb/1.0.0/



Database Open Access

MIT-BIH Arrhythmia Database

George Moody, Roger Mark

Published: Feb. 24, 2005. Version: 1.0.0

MIT-BIH Arrhythmia Database expanded *(Feb. 24, 2005, midnight)*

The entire MIT-BIH Arrhythmia Database is now freely available on PhysioNet. Somewhat more than half of the database has been available here since PhysioNet's inception; the remainder has now been posted.

When using this resource, please cite the original publication:
Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209)

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Share

Access

Access Policy:
Anyone can access the files, as long as they conform to the terms of the specified license.

License (for files):
Open Data Commons Attribution License v1.0

Background

Since 1975, our laboratories at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center) and at MIT have supported our own research into arrhythmia analysis and related subjects. One of the first major products of that effort was the MIT-BIH Arrhythmia Database, which we completed and began distributing in 1980. The database was the first generally available set of standard test material for evaluation of arrhythmia detectors, and has been used for that purpose as well as for basic research into cardiac dynamics at more than 500 sites worldwide. Originally, we distributed the database on 9-track half-inch digital tape at 800 and 1600 bpi, and on quarter-inch IRIG-format FM analog tape. In August, 1989, we produced a CD-ROM version of the database.

103