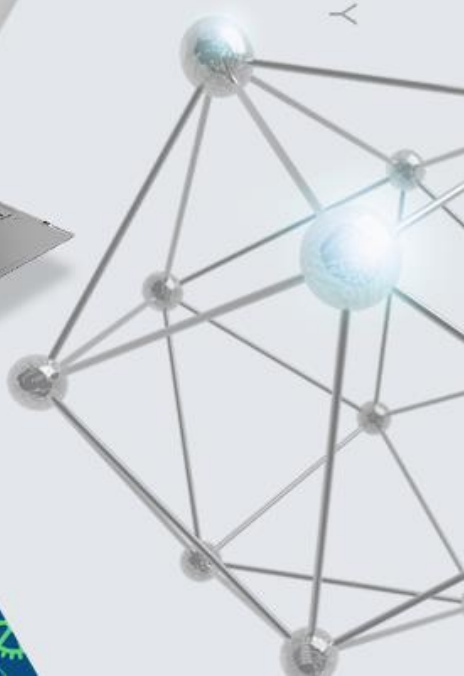
The 4th Industrial Revolution is characterized by super connectivity and super intelligence, where various products and services are connected to the network, and artificial intelligence and information communication technologies are used in 3D printing, unmanned transportation, robotics, Of the world's most advanced technologies.

TECHNOLOGY

인공지능 자연어 처리

인공지능 자연어 처리 절차 알아보기



The 4th Industrial Revolution is characterized by super connectivity and super intelligence, where various products and services are connected to the network, and artificial intelligence and information communication technologies are used in 3D printing, unmanned transportation, robotics, Of the world's most advanced technologies.

인공지능 자연어 처리 절차 알아보기

학/습/목/표

1. 인공지능 딥러닝 모델과 학습 절차에 대해 설명할 수 있다.
2. 텍스트 전처리와 네트워크 모델 설계를 설명할 수 있다.
3. 모델을 훈련하고 예측 및 평가하는 과정을 설명할 수 있다.

학/습/내/용

1. 인공지능 딥러닝 모델과 학습 절차
2. 텍스트 전처리와 네트워크 모델 설계
3. 모델 훈련과 예측 및 평가

1. 인공지능 딥러닝 모델과 학습 절차

1) 인공지능 딥러닝 모델

(1) 인공지능 딥러닝 모델이란?

- 관측 값(Observations)을 축적하여 데이터 구축
- 머신러닝/딥러닝 알고리즘으로 데이터를 학습시켜 생성된 일반화된 규칙

(2) 인공지능 딥러닝 모델 구축

- 머신러닝/딥러닝 알고리즘을 이용하여 데이터와 해답(레이블)을 학습시켜 일반화된 규칙인 모델(Model) 구축

데이터
해답(레이블)

Distance	Time	Distance	Time
4.9m	1s	44.1m	3s
19.6m	2s	78.5m	4s

머신러닝/딥러닝 알고리즘

일반화된 규칙 모델

$g = 9.8m/s^2$

1.인공지능딥러닝모델과학습절차

2) 인공지능 딥러닝 학습

(1) 가중치(Weight)

- 입력신호의 강도를 표현($W1, W2 \cdots, Wn$)

(2) 입력신호의 총합(Summation)

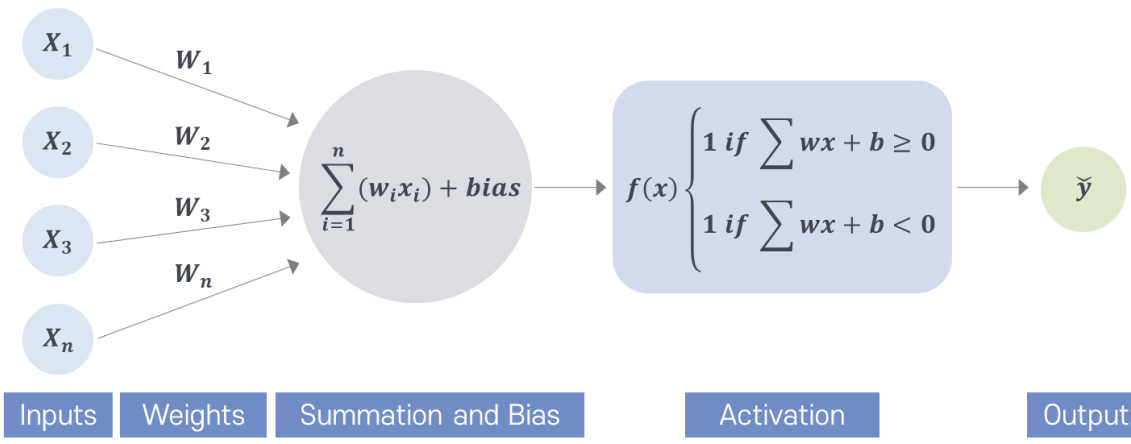
- 각 입력신호에 가중치를 곱하여 합한 값
  ($W1 \cdot x1 + W2 \cdot x2 + \cdots + Wn \cdot xn = \Sigma Wi \cdot xi$)

(3) 활성화 함수(Activation Function)

- 신호의 총합을 출력신호로 변환

(4) 인공지능 딥러닝 학습은?

- 인공신경망에서 최적의 가중치($W1, W2 \cdots, Wn$) 값들을 찾아 인공지능 딥러닝 모델을 구축하는 과정



1.인공지능딥러닝모델과학습절차

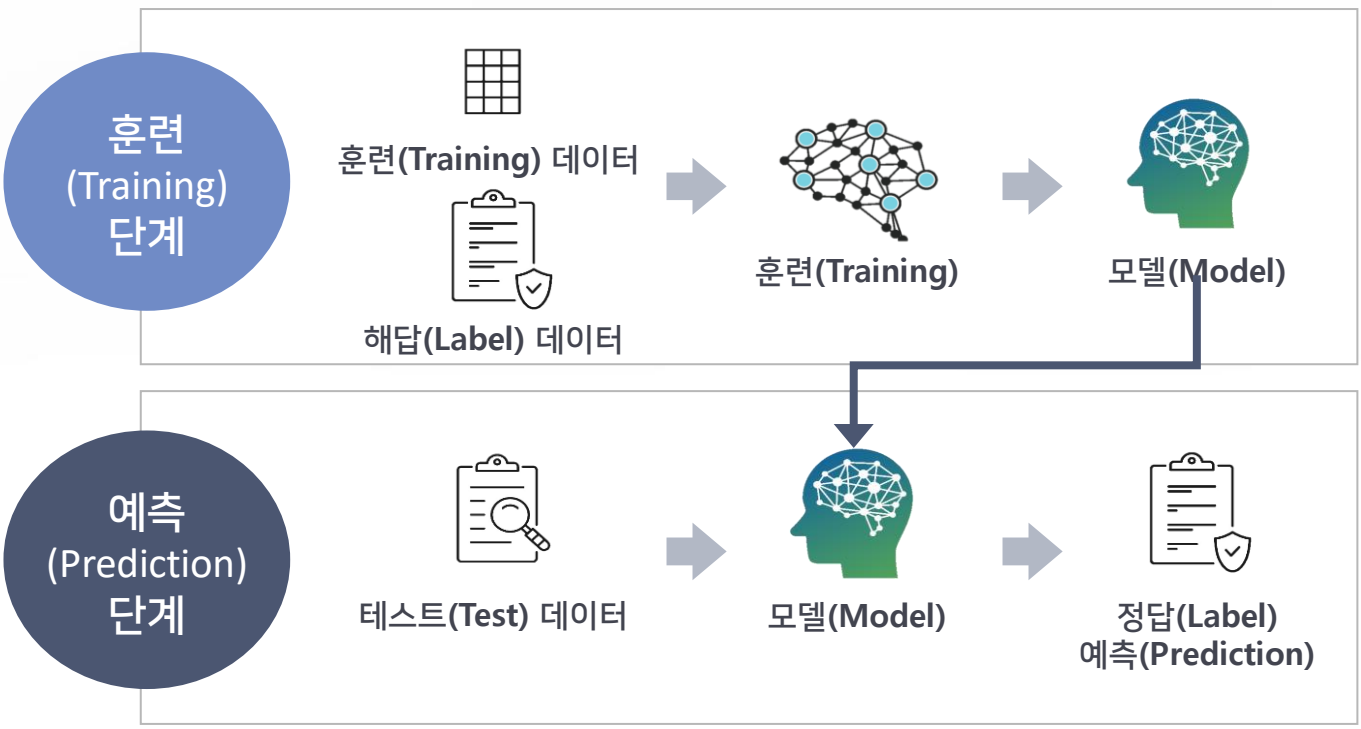
3) 인공지능 딥러닝 학습단계와 예측단계

(1) 훈련(Training) 단계

- Training 데이터와 Label 데이터를 설계된 모델에 입력하여 학습을 진행하고 결과물로 모델을 생성

(2) 예측(Prediction) 단계

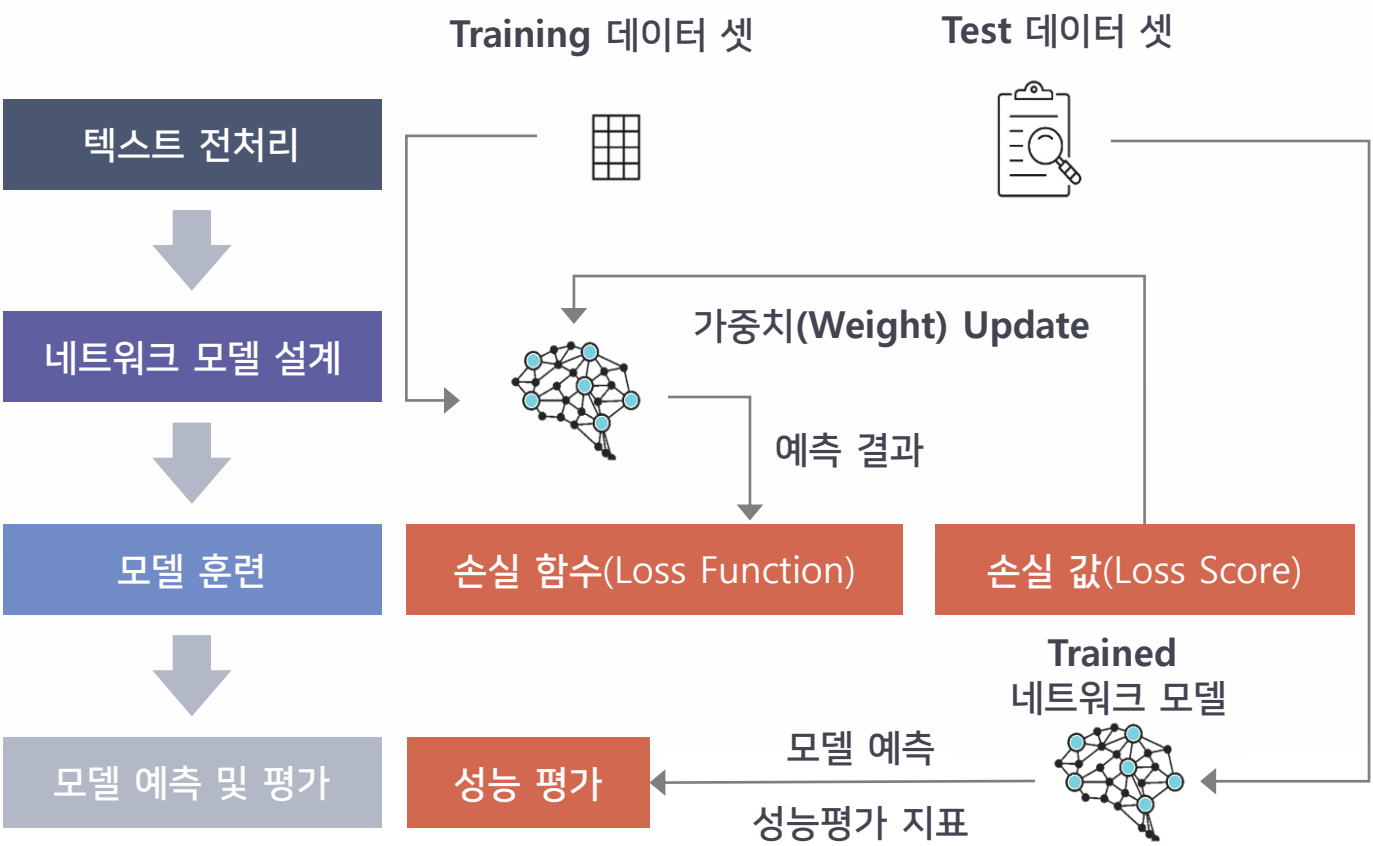
- 훈련 단계에서 생성된 모델에 Test 데이터를 입력하여 결과를 예측



1.인공지능딥러닝모델과학습절차

4) 인공지능 딥러닝 자연어 처리 학습 절차

딥러닝 자연어 처리 학습 절차는 텍스트 전처리, 네트워크 모델 설계, 모델 훈련, 모델 예측 및 평가 절차에 따라 진행함



1.인공지능딥러닝모델과학습절차

5) 인공지능 딥러닝 훈련 사이클

(1) Y′ 예측 및 Y′와 Y의 차이 값 계산

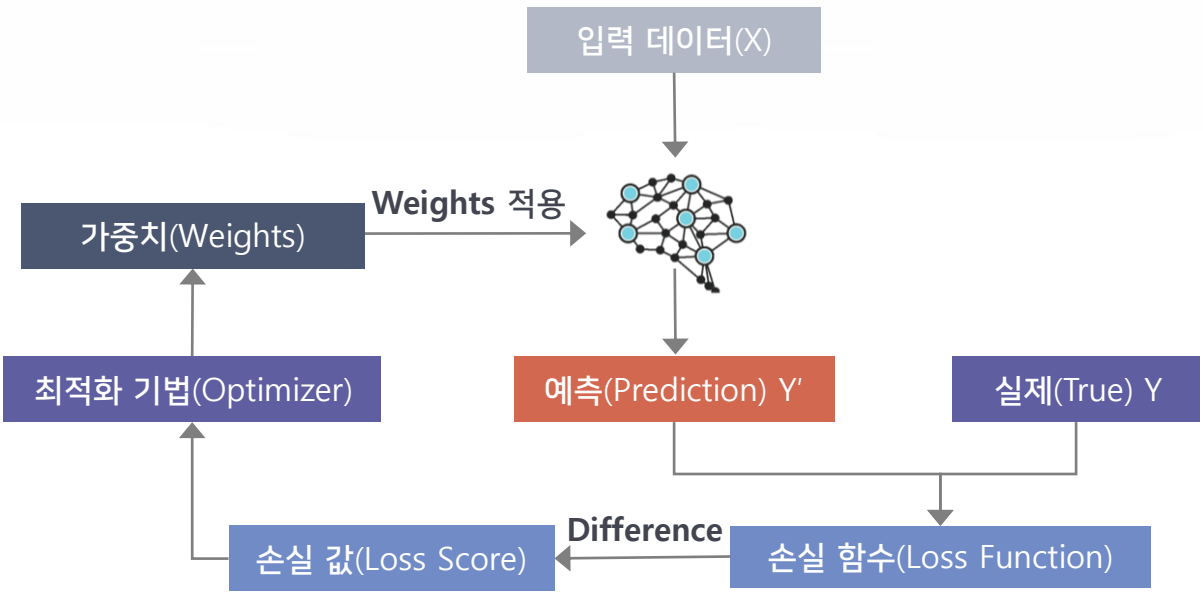
- 입력된 데이터(X)에 대해 뉴럴 네트워크를 통해 정답 후보 값인 Y′를 예측하고, 실제 정답인 Y와 비교하여 차이 계산

(2) 가중치(Weights) 계산

- 계산된 차이 만큼의 손실 값(Loss Score)을 활용하여 최적화 기법으로 새롭게 업데이트할 가중치(Weights) 계산

(3) Y′ 예측

- 가중치를 업데이트하여 입력 데이터 X에 대한 새로운 정답 후보 값 Y′ 예측

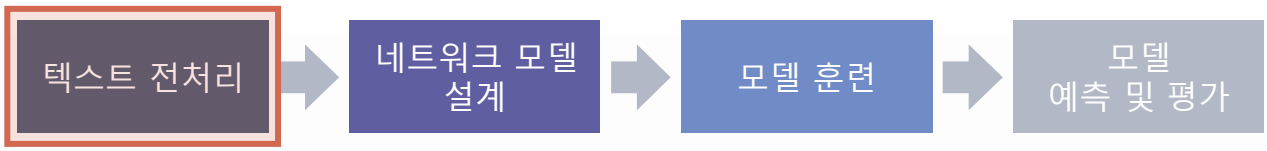


2 텍스트 전처리와 네트워크 모델 설계

1) 텍스트 전처리 개요

(1) 코퍼스(Corpus)란?

- 말뭉치라고도 하며 여러 문장들로 구성된 문서의 집합



(2) 텍스트 전처리(Preprocessing) 단계는?

- 정제, 불용어 제거, 어간 추출, 토큰화 및 문서 표현 등의 작업을 수행하는 단계
- 파이썬의 torchtext는 텍스트 전처리를 위한 파이썬 라이브러리

2 텍스트 전처리와 네트워크 모델 설계

1) 텍스트 전처리 개요

- 정제(Cleaning) : 특수문자 등과 같은 불필요한 노이즈 텍스트 제거 및 대소문자 통일
  - 특수문자 예시: '!"#$%&₩'()*+,-./:;<=>?@[₩₩]^_`{|}~'
  - 대소문자 통일 예시: korea, Korea → KOREA

- 불용어 제거(Stop word Elimination) : 전치사, 관사 등 문장이나 문서의 특징을 표현하는데 불필요한 단어를 제거하는 단계

- 어간 추출 : 단어의 기본 형태를 추출하는 단계
  - 어간 추출 예시: stem, stems, stemmed, stemmer → stem

- 토큰화 : 코퍼스(Corpus)에서 분리자(Separator)를 포함하지 않는 연속적인 문자열 단위로 분리
  - 토큰화 예시: 한글 토큰화 결과입니다. →['한글', '토큰', '화', '결과', '입니다', '.']

☞ 참고: 파이썬에서 영문 토큰화는 nltk를 사용, 한글 토큰화는 konlpy를 사용

- 문서 표현 : 주어진 문서나 문장을 하나의 벡터로 표현하는 단계
  -단어들을 모두 인덱싱(Indexing)하고
  주어진 문서에 존재하는 단어의 빈도수를 사용하여 문서를 표현
  - 문서표현 방법: One-hot 인코딩, Word2Vec 등

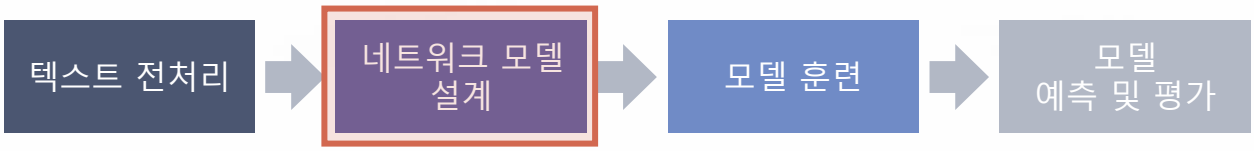
1) 텍스트 전처리 개요

(3) One-hot 인코딩(Encoding)이란?

- 전체 레이블(Label) 수에 대해 표현하고자 하는 특정 단어(Word)만 '1'로 활성화한 벡터로 변환하는 방법
- 단어의 의미와 관계를 전혀 고려하지 않음
- 벡터의 크기는 총 단어 수만큼의 Sparse Vector

사과	1	0	0	0	0
개	0	1	0	0	0
배	0	0	1	0	0
고양이	0	0	0	1	0
토마토	0	0	0	0	1

2) 네트워크 모델 설계

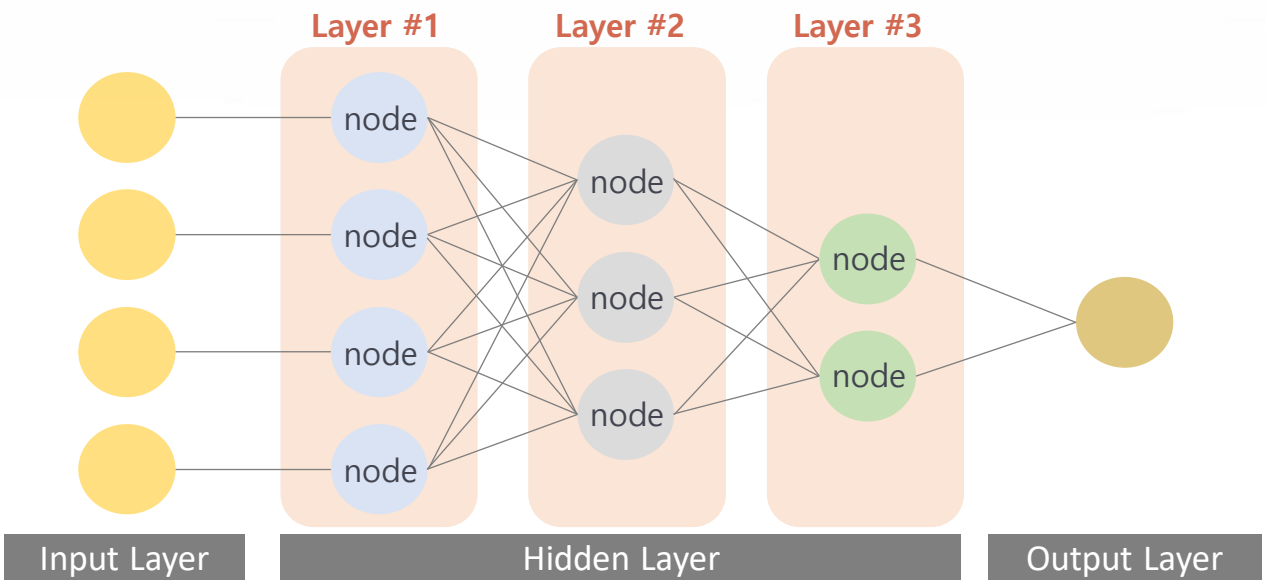


(1) 뉴럴 네트워크의 구조는?

- 여러 은닉층(Hidden Layer)을 가지는 모델
- 하나의 은닉층은 여러 노드로 구성됨

(2) 네트워크 모델 설계란?

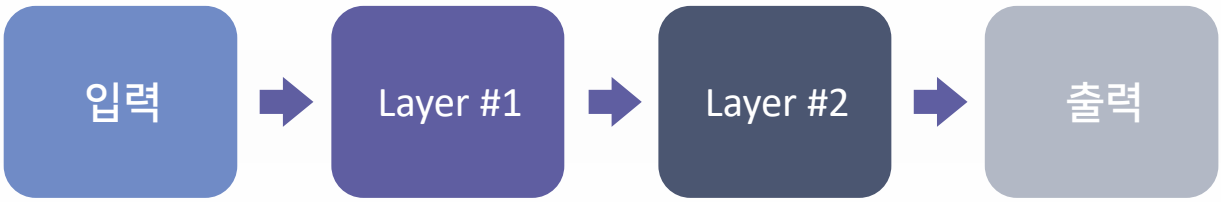
- 인공신경망(Artifitial Neural Network)은 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)으로 구성



2) 네트워크 모델 설계

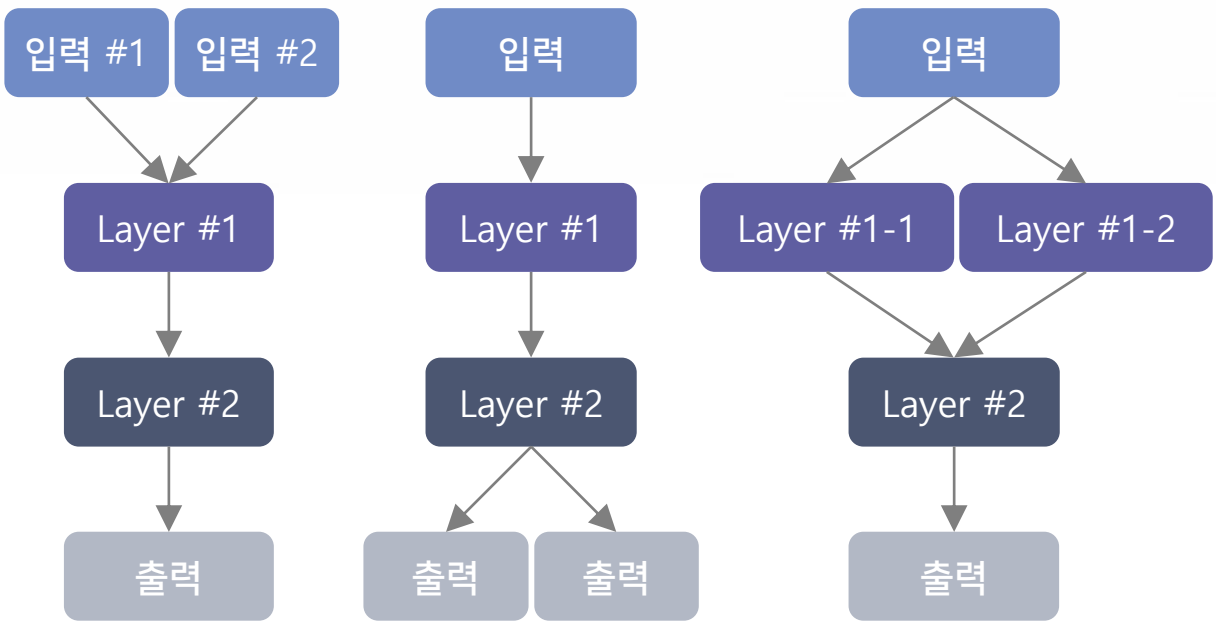
(3) Sequencial 구조란?

- 단일 흐름의 순차적 모델 구성에 사용되는 구조



(4) Functional API 구조란?

- 다중 입출력 및 분기 흐름의 모델 구성에 사용되는 구조



3. 모델 훈련과 예측 및 평가

1) 모델 훈련

텍스트 전처리 → 네트워크 모델 설계 → **모델 훈련** → 모델 예측 및 평가

(1) 모델 훈련(Training) 단계는?

- 설계한 뉴럴 네트워크를 훈련 데이터(Training Data)를 사용하여 가중치를 학습하며 훈련을 진행하는 단계
- 훈련이 종료되면 새로운 모델이 생성

(2) 딥러닝 학습 데이터

- 딥러닝 학습데이터는 훈련(Training) 데이터, 검증(Validation) 데이터 및 테스트(Test) 데이터로 분할(Split)하여 사용

전체 데이터(Original Data) 100%

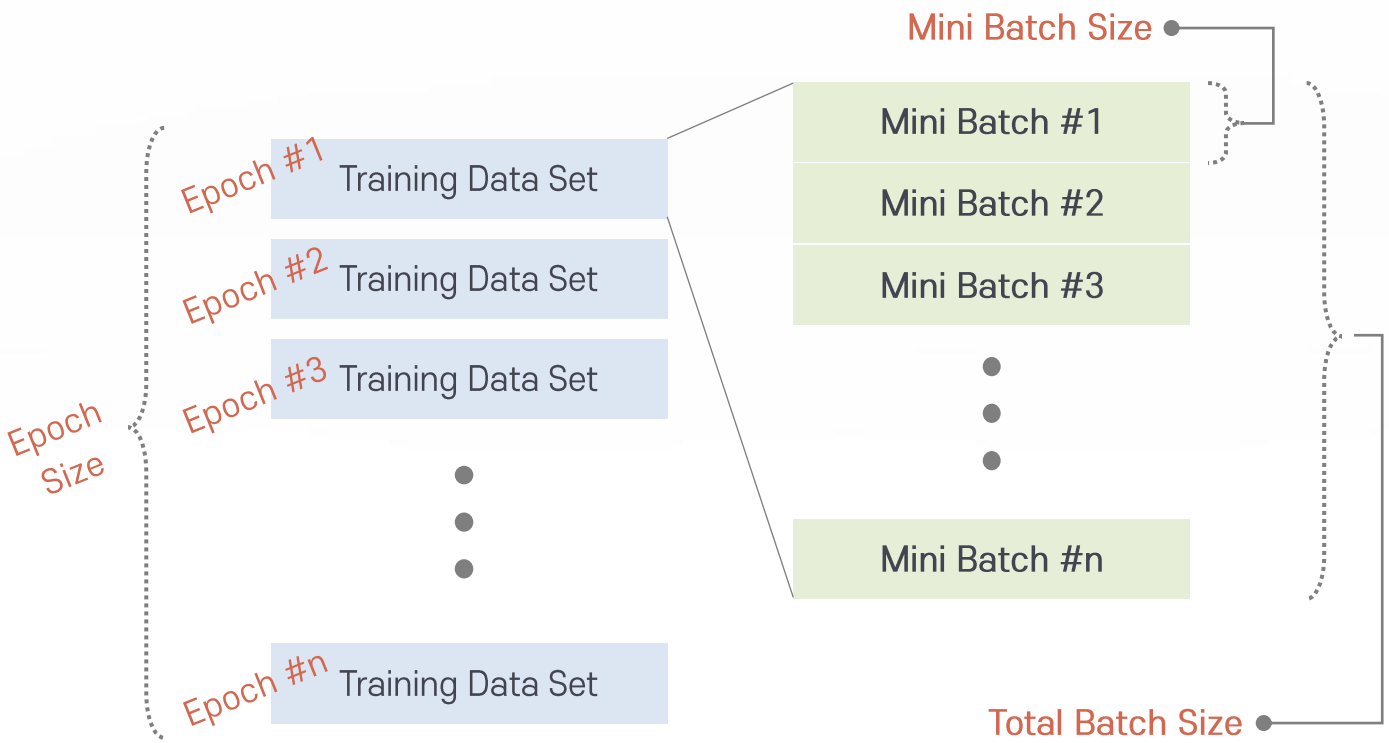
훈련(Training) 데이터		테스트(Test) 데이터
85%, 80%, 70%		15%, 20%, 30%
훈련(Training) 데이터	검증(Validation) 데이터	테스트(Test) 데이터
50%	30%	15%, 20%, 30%
모델 구축을 위한 데이터(문제집)	모델 성능향상을 위한 데이터(모의고사)	예측 및 성능 평가를 위한 데이터(수능시험)

3. 모델 훈련과 예측 및 평가

1) 모델 훈련

(3) 딥러닝의 반복 훈련

- 딥러닝 반복 학습은 동일한 Train 데이터 셋을 Mini Batch 크기 만큼 분할하여 학습
- Mini Batch Size : 가중치(Weights)를 한 번 업데이트하는 주기
- Epoch : Training Data Set 입력 데이터로 모두 한 번 사용한 주기
- 전체 Epoch Size는 동일한 Training Data Set으로 반복 학습한 수



3. 모델 훈련과 예측 및 평가

2) 모델 예측 및 평가

텍스트 전처리 → 네트워크 모델 설계 → 모델 훈련 → **모델 예측 및 평가**

(1) 모델 예측 및 평가 단계는?

- 테스트 데이터(Test Data)를 사용하여 예측을 수행하고 모델의 성능을 평가하는 단계
- 모델의 성능 평가지표는 일반적으로 정확도(Accuracy)를 많이 사용

(2) Keras 프레임워크의 모델 예측 및 평가 예시

```
test_loss, test_acc = model.evaluate(test_images, test_labels)
```

10000/10000 [= = = = = = = = = =] – 0s 30us/step

```
print ( ‘ test_acc: ’ , test_acc)
```

test_acc: 0.9795

Test 데이터를 이용한 모델의 정확도는 97.95%

Summary

1. 인공지능 딥러닝 모델과 학습 절차

1) 인공지능 딥러닝 모델

- 관측 값(Observations)을 축적하여 데이터 구축
- 머신러닝/딥러닝 알고리즘으로 데이터를 학습시켜 생성된 일반화된 규칙

2) 딥러닝 학습

- 최적의 가중치(Weight) 값들을 찾는 과정

3) 딥러닝 학습 절차

- 텍스트 전처리, 네트워크 모델 설계, 모델 훈련, 모델 예측 및 평가 절차에 따라 진행

4) 인공지능 딥러닝 훈련(Training)단계와 예측(Prediction) 단계

- 인공지능 딥러닝 훈련(Training)단계에서는 Training 데이터와 Label 데이터를 설계된 모델에 입력하여 학습을 진행하고 결과물로 모델을 생성
- 인공지능 딥러닝 예측(Prediction) 단계에서는 훈련 단계에서 생성된 모델에 Test 데이터를 입력하여 결과를 예측

Summary

2. 텍스트 전처리와 네트워크 모델 설계

1) 텍스트 전처리

- 정제, 불용어 제거, 어간 추출, 토큰화 및 문서 표현 등의 작업을 수행하는 단계

2) One-hot 인코딩(Encoding)

- 전체 레이블(Label) 수에 대해 표현하고자 하는 특정 단어(Word)만 '1'로 활성화한 벡터로 변환하는 방법으로 단어의 의미와 관계를 전혀 고려하지 않음

3) 인공신경망

- 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)으로 구성

3. 모델 훈련과 예측 및 평가

- 딥러닝 학습 데이터 중 훈련 데이터, 검증 데이터 및 테스트 데이터의 비율은 각각 50% - 30% - 20%로 이루어 짐
- 모델의 성능 평가지표로 일반적으로 많이 사용하는 것은 정확도(Accuracy) 지표