핵심 기술 조사 및 실험

MUD

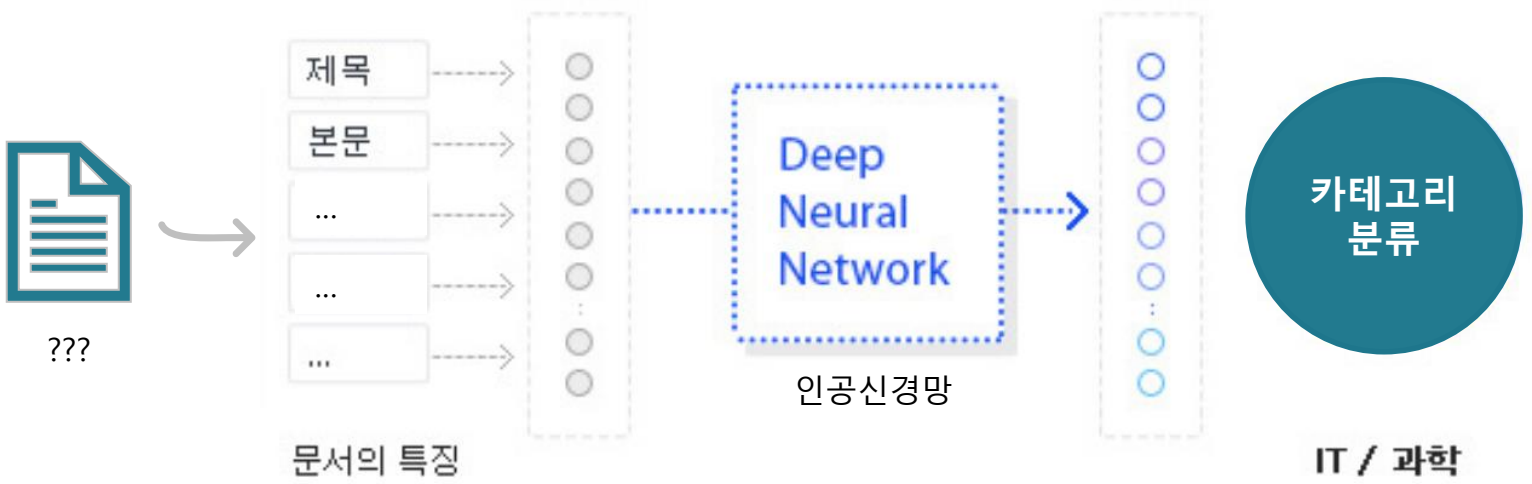
Meeting Using Deep Learning

201600599
김아연
2019.09.24

CONTENTS

1.Part 소개

원시 텍스트 문서 주제 분류



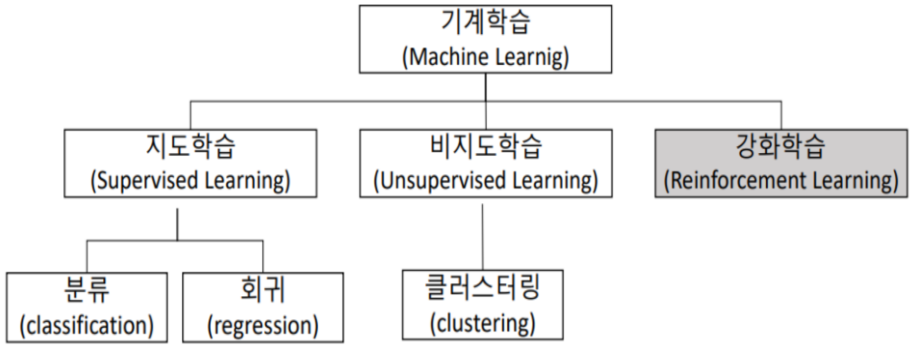
<그림1>

- 딥러닝을 이용하여 **문서의 주제를 자동으로 분류**하는 시스템을 구현합니다.
- STT를 통해 생성된 하나의 원시 텍스트에는 **한 가지 이상의 주제**가 담겨있습니다.
- 문서 내 담겨있는 주제의 개수를 몰라도 자동으로 **여러 주제로 분류**해줍니다.

2. 기술 소개

머신러닝과 딥러닝

▪ 머신러닝

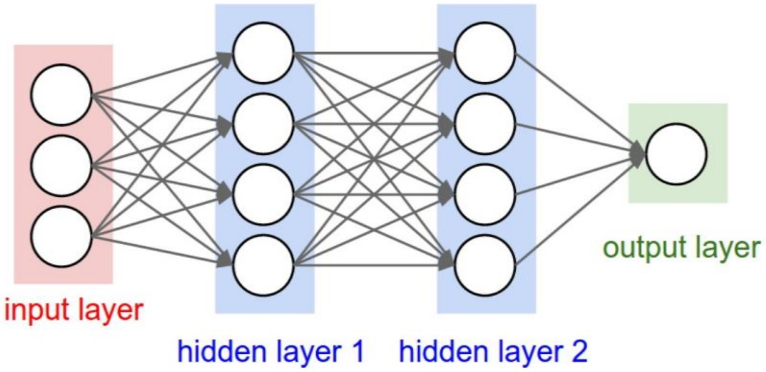


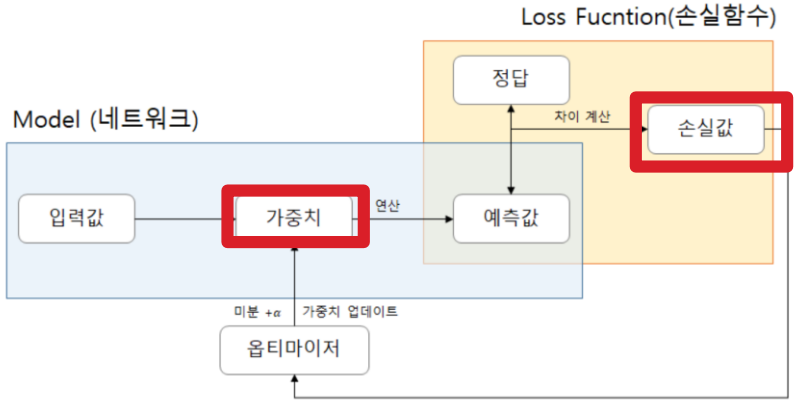
- 프로그래머가 룰을 정해주던 기존의 방식과 달리 **컴퓨터가 직접 룰을 학습**합니다.
- 즉 , 입력 데이터를 **의미 있는 데이터로 변환, 유용한 표현을 학습**하는 것입니다.
- 사람의 감독하에 훈련하는 것인지 그렇지 않은 것인지에 따라 지도학습과 비지도학습으로 구분됩니다.

2. 기술 소개

머신러닝과 딥러닝

▪ 딥러닝

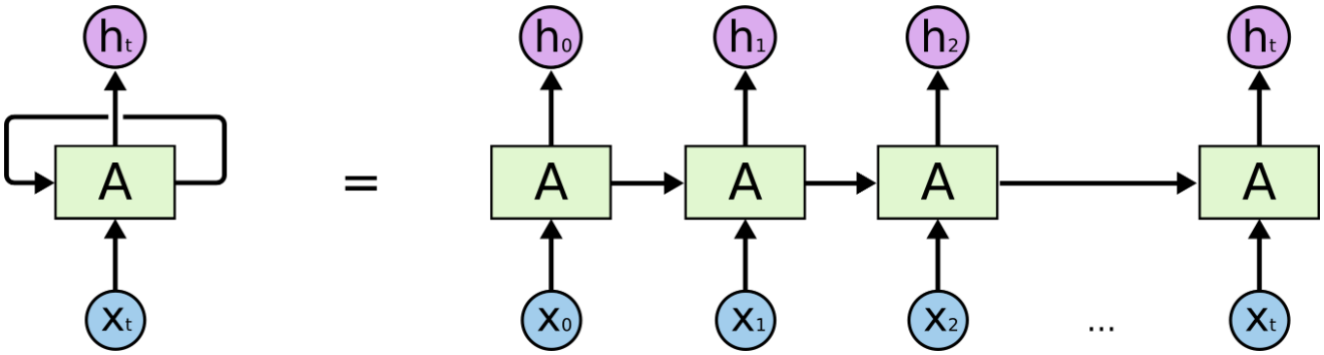




- 딥러닝은 연속된 **층(layer)**을 쌓아 올려 **신경망**이라는 모델을 사용해 학습합니다.
- **학습**은 입력을 타겟에 매핑하기 위해 모든 층의 **가중치 값을 찾는 것**입니다.
- 예측(Y')과 실제 타겟(Y)의 차이를 측정하기 위해 **손실함수**를 사용합니다.
- **훈련 반복을** 통해 **손실값을 최소화**하는 것을 목적으로 한다.

2. 기술 소개

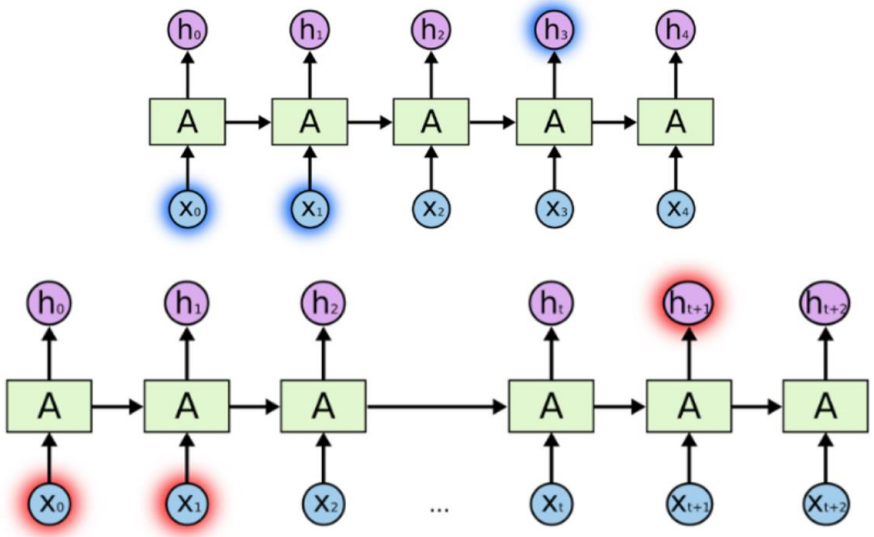
■ RNN(Recurrent Neural Network)



- 입력과 출력을 **시퀀스 단위**로 처리하는 모델입니다.
- 이전의 데이터가 그 다음 데이터 출력에 영향을 줍니다.
- 음성인식, 언어 모델링, 번역, 이미지 캡셔닝 등 여러 분야에서 성과를 내고 있습니다.
- 바닐라 RNN은 비교적 짧은 시퀀스(sequence)에 대해서만 효과를 보이는 단점이 있습니다.

2. 기술 소개

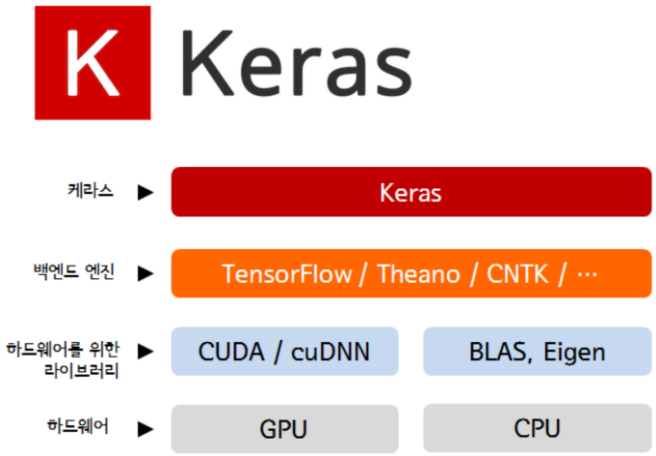
▪ LSTM(Long Short-Term Memory)



- RNN의 **장기 의존성 문제를 해결**합니다.
- RNN의 hidden state에 **cell state**를 추가한 구조입니다.
- LSTM은 은닉층의 메모리 셀에 **입력 게이트, 망각 게이트, 출력 게이트**를 추가하여 불필요한 기억을 지우고, 기억해야할 것들을 정합니다.

3. 기술 실험

▪ 케라스



<작업 흐름>

1. 입력 텐서와 타깃 텐서로 이루어진 **훈련 데이터를 정의**
2. 입력과 타깃을 매핑하는 층으로 이루어진 **네트워크(모델)**를 정의
3. 손실 함수, 옵티마이저, 모니터링하기 위한 **측정 지표를 선택**하여 학습 과정을 설정
4. 훈련 데이터에 대해 모델의 fit() 메서드를 **반복적으로 호출**

- 케라스는 딥러닝 모델을 간편하게 만들고 훈련시킬 수 있는 파이썬을 위한 **딥러닝 프레임워크**이다.
- 사용하기 쉬운 API를 가지고 있어 딥러닝 모델의 프로토타입을 빠르게 만들 수 있다.
- 케라스로 작성한 코드는 아무런 변경 없이 언제든지 백엔드를 바꿀 수 있다.

3. 기술 실험

▪ 예제 - 로이터 뉴스 분류

1. 데이터 준비: <로이터 뉴스>

1.1 훈련용 뉴스 기사: 8982 / 테스트용 뉴스 기사: 2246/ 카테고리: 46
1.2 단어 사용, 문서 길이 지정: max_word = 1000, max_len = 100

2. LSTM모델 생성

```
model = Sequential()
model.add(Embedding(1000, 120))
model.add(LSTM(120))
model.add(Dense(46, activation='softmax'))
```

3. 모델 컴파일

```
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

4. 모델 반복 학습

```
history = model.fit(X_train, y_train, batch_size=100, epochs=20, validation_data=(X_test,y_test))
```

```
Epoch 20/20
8982/8982 [==============================] - 10s 1ms/step - loss: 0.7941 - acc: 0.7968 - val_loss: 1.2014 - val_ac
c: 0.7035
```

3. 기술 실험

▪ 적용

1. 사용한 데이터: <네이버지식백과>

1.1 훈련용 : 80000 / 테스트용 : 20000/ 카테고리: 5

	타겟	1	2	입력	4
0	기술공학	기계자동차금속	백 도어 내장 공구 박스	백 도어 설치 장통 말 공구 진열 방식 수납 때문 미관 사용 상의 장점	0
1	기술공학	컴퓨터통신IT	재핑 효과 [zapping effect, -效果]	채널 중간 채널 시청률 현상 재핑 이란 방송 프로그램 시작 전후 노출 광고 피 위해...	0
2	문화예술	미술	경매 외에는 어떤 미술시장이 있나	미술관 달리 화랑 회사 그림 위 목적 전시 곳 돈 주머니 미술품 쇼핑 위해 화랑 매...	1
3	문화예술	음악	약 [籥]	약 단소 세로 취구 입김 불어 구멍 개 공 손가락 여 구멍 개 예종 수용 아악 통해...	1
4	인문과학	인문과학 일반	청유형 문장	말 이 이 행동 요청 문장 말 문장 끝 종결 어미 자기 생각 느낌 이 여러 가지 방...	2

1.2 입력 데이터
- 명사만 사용
- 불용어 지정: 것,이,그,수,등....

1.3 단어 사용, 문서 길이 지정
max_word = 5000, max_len = 500

3. 기술 실험

■ 적용

3. 결과

```
print("정확도 : %.4f" % (model.evaluate(X_test, y_test)[1]))
```

```
20000/20000 [==============================] - 25s 1ms/step
정확도 : 0.8083
```

```
#L_category_num = {'기술공학': 0, '문화예술': 1, '인문과학': 2, '역사문화': 3, '사회과학': 4}
```

```
str = ['안드로이드 스마트폰 갤럭시 새롭게 탄생 아이폰 쓰레기 애플 컴퓨터 보안 장비 보안 이슈 블록체인 안드로이드 스마트폰 성능 안드로이드 스마트폰 갤럭시 새롭게 탄생
```

```
Out[495]: array([0], dtype=int64)
```

기술공학

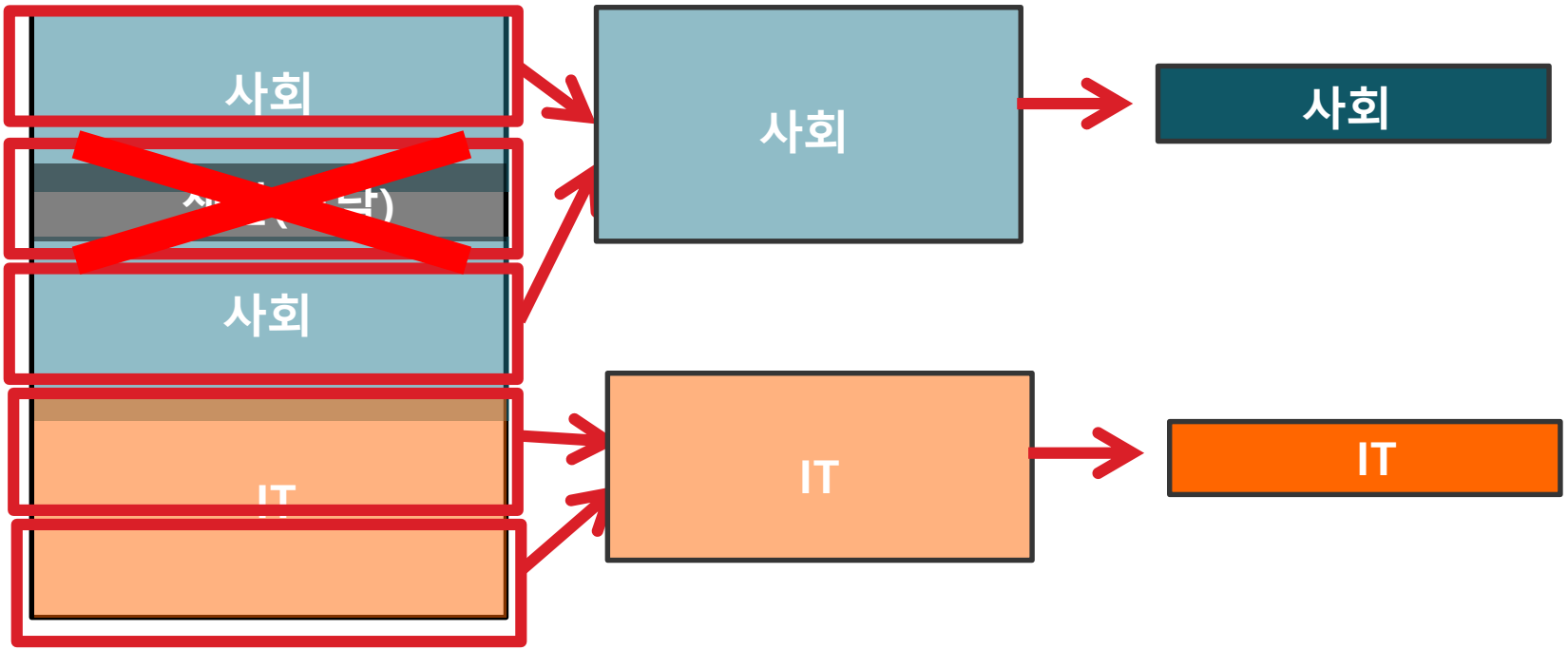
```
str = ['''조선시대 일제강점기 한국사 역사 고구려 신라 백제
''']
```

```
Out[506]: array([3], dtype=int64)
```

역사문화

4. 향후 계획

- **적용 과정**



Raw text　　주제 분류　　요약

4. 향후 계획

▪ 고려할 점

1. 좋은 품질의 데이터셋 구하기
- 회의 데이터는 구하기가 쉽지 않다. 카테고리 분류가 많은 지식백과를 크롤링하여 데이터로 사용하려 하였지만 데이터가 좋지 않다고 판단하였다.

2. 문단 단위 지정
- 회의는 구어체이다 보니 문단 지정이 되어있지 않다. 어떠한 기준으로 문단을 나누어야 좋은 결과가 나올지 실험을 계속 해보아야한다.

3. 학습 전 데이터 전처리
- 형태소 분석기는 어떤 것을 사용할 지, 단어는 어떤 품사를 사용할 지, 불용어 지정을 어떻게 해야할 지 생각해보아야한다.

4. 추가 실험을 통해 더 좋은 신경망 모델 생성
- 더 크거나 작은 층을 사용하여 보며 보다 좋은 성능을 내는 모델을 생성한다.

5. 출처 및 참고 문헌

출처

	항목	출처
1	그림1	https://devblog.zum.com/279
2	모두의 딥러닝	https://thebook.io/006958/
3	케라스 창시자에게 배우는 딥러닝	https://thebook.io/006975/
4	모두를 위한 머신러닝/딥러닝 강의	http://hunkim.github.io/ml/
5		
6		
7		
9		
10		

감사합니다