

시계열 데이터와 순환 신경망

Preview

■ 시계열 데이터

- 시간 정보가 들어 있는 데이터
 - 예) 문장 "세상에는 시계열 데이터가 참 많다"
 - 단어가 나타나는 순서가 중요
 - 샘플의 길이가 다름

■ 시계열 데이터를 인식하는 고전적인 모델

- ARIMA(autoregressive integrated moving average)와 SARIMA(seasonal ARIMA)
- Prophet 등

■ 시계열 데이터를 인식하는 딥러닝 모델

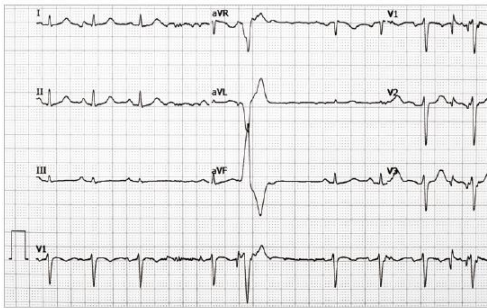
- 순환 신경망 recurrent neural network(RNN)
- LSTM_{long short-term memory}: 선별 기억 능력을 갖춰 장기 문맥 처리에 유리

8.1 시계열 데이터의 이해

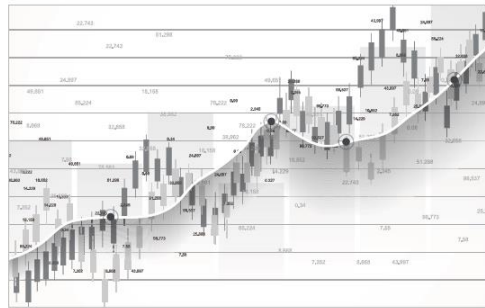
■ 시계열 데이터

- 시간 축을 따라 신호가 변하는 동적 데이터
- SVM, 다층 퍼셉트론, 깊은 다층 퍼셉트론, 컨볼루션 신경망
 - 정적 데이터를 한꺼번에 입력받기 때문에 시계열 데이터에 부적합
 - 시계열 데이터를 정적 데이터로 변환하여 입력하면 정보 손실이 큼

■ 딥러닝에서는 시계열 특성을 반영하는 순환 신경망 또는 LSTM 활용



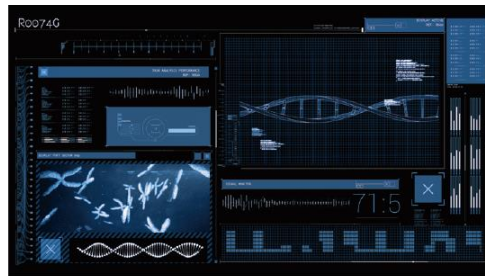
(a) 심전도



(b) 주식 시세



(c) 음성 인식 데이터



(d) 유전자 염기 서열

그림 8-1 다양한 시계열 데이터

8.1.1 시계열 데이터의 특성

■ 시계열 데이터의 독특한 특성

- 요소의 순서가 중요
 - 예) "세상에는 시계열 데이터가 참 많다" 를 "시계열 참 데이터가 많다 세상에는" 으로 바꾸면 의미 훼손
- 샘플의 길이가 다름
 - 예) 짧은 발음 "인공지능" 과 긴 발음 "인~공~~지~능"
- 문맥 의존성
 - 예) "시계열은 앞에서 말한 바와 ... 특성이 있다" 에서 "시계열은" 과 "특성이 있다" 는 밀접한 관련성
- 계절성
 - 예) 상추 판매량, 미세먼지 수치, 항공권 판매량 등

■ 시계열 데이터의 표현 $\mathbf{x} = (\mathbf{a}^1 \ \mathbf{a}^2 \cdots \mathbf{a}^t)$ (8.1)

- 가변 길이로 벡터의 벡터임
- 예) 매일 기온, 습도, 미세먼지 농도를 기록한다면, $\mathbf{a}^1 = (23.5, 42, 0.1)$, $\mathbf{a}^2 = (25.5, 45, 0.08)$, ...

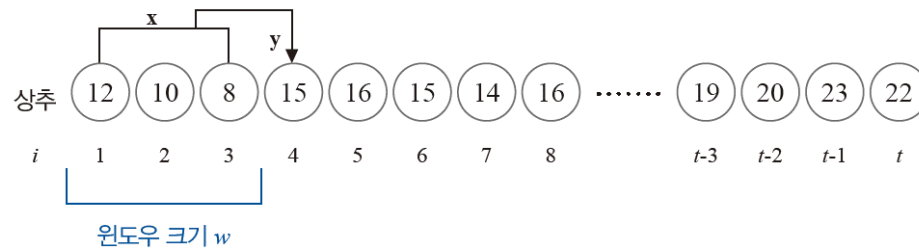
8.1.2 미래 예측을 위한 데이터 준비

- 순환 신경망은 유연한 구조라 여러 문제에 적용 가능
 - 대표적 응용은 미래 예측(prediction 또는 forecasting)
 - 내일 주가 예측
 - 내일 날씨 예측
 - 기계의 고장 예측
 - 풍속과 풍향 예측(풍력 발전기의 효율 향상)
 - 농산물 가격/수요량 예측 등
 - 언어 번역에 응용
 - 음성 인식에 응용
 - 생성 모델에 응용(예, 사진을 보고 설명 문장 생성)

8.1.2 미래 예측을 위한 데이터 준비

■ 예, 농산물 수요량 예측 문제에서 데이터

- 농산물 유통업자가 5년 동안 매일 판매량을 기록했다면 길이 $t=365*5=1825$ 인 샘플
- 하나의 긴 샘플을 가지고 어떻게 모델링하고 어떻게 미래를 예측하나?
 - 윈도우 크기(w) 단위로 패턴을 잘라 여러 개의 샘플을 수집([그림 8-2]에서는 $w=3$)
 - 얼마나 먼 미래를 예측할 지 지정하는 수평선 계수 h ([그림 8-2]에서는 $h=1$)



(a) 입력 데이터 샘플링

샘플	x	y
1	(12, 10, 8)	15
2	(10, 8, 15)	16
3	(8, 15, 16)	15
4	(15, 16, 15)	14
...
$t-w$	(19, 20, 23)	22

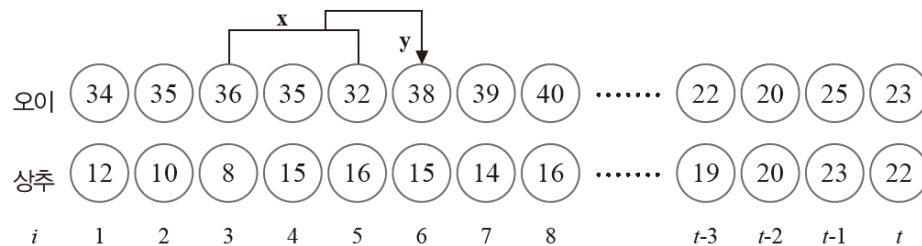
(b) 입력 패턴에서 생성한 샘플

그림 8-2 미래 예측 문제에서 샘플 생성하기(단일 채널)

8.1.2 미래 예측을 위한 데이터 준비

■ 다중 품목을 표현하는 데이터

- 벡터의 벡터 구조
- 예) 상추와 오이라는 두 품목을 동시에 고려하는 데이터



(a) 입력 패턴

샘플	x	y
1	$((34, 12), (35, 10), (36, 8))$	$(35, 15)$
2	$((35, 10), (36, 8), (35, 15))$	$(32, 16)$
3	$((36, 8), (35, 15), (32, 16))$	$(38, 15)$
4	$((35, 15), (32, 16), (38, 15))$	$(39, 14)$
...
$t - w$	$((22, 19), (20, 20), (25, 23))$	$(23, 22)$

(b) 입력 패턴에서 생성한 샘플

그림 8-3 미래 예측 문제에서 샘플 생성하기(다중 채널)

8.1.3 시계열 데이터 사례: 비트코인 가격

■ 코인데스크에서 데이터 다운로드

- www.coindesk.com

[Currency, Date, Closing Price(USD),
24h Open(USD), 24h High(USD), 24h Low(USD)]의
6개 열로 구성



	A	B	C	D	E	F	G
1	Currency	Date	Closing Price	24h Open	24h High	24h Low (USD)	
2	BTC	2019-02-28	3772.94	3796.64	3824.17	3666.52	
3	BTC	2019-03-01	3799.68	3773.44	3879.23	3753.8	
4	BTC	2019-03-02	3811.61	3799.37	3840.04	3788.92	
5	BTC	2019-03-03	3804.42	3806.69	3819.19	3759.41	
6	BTC	2019-03-04	3782.66	3807.85	3818.7	3766.24	
7	BTC	2019-03-05	3689.86	3783.36	3804.35	3663.48	
8	BTC	2019-03-06	3832.08	3701.05	3866.72	3688.7	
9	BTC	2019-03-07	3848.96	3832.59	3881.97	3802.52	
10	BTC	2019-03-08	3859.84	3848.96	3890.75	3827.67	
11	BTC	2019-03-09	3828.37	3859.8	3918	3778.52	
12	BTC	2019-03-10	3898.87	3841.89	3948.88	3832.14	
13	BTC	2019-03-11	3899.66	3916.76	3921.93	3865.92	
14	BTC	2019-03-12	3851.25	3899.46	3913.46	3819.93	

(a) 코인데스크에서 데이터를 다운로드하기

(b) 비트코인 가격이 저장된 데이터 프레임

그림 8-4 코인데스크 사이트에서 다운로드한 비트코인 가격 데이터

8.2 순환 신경망

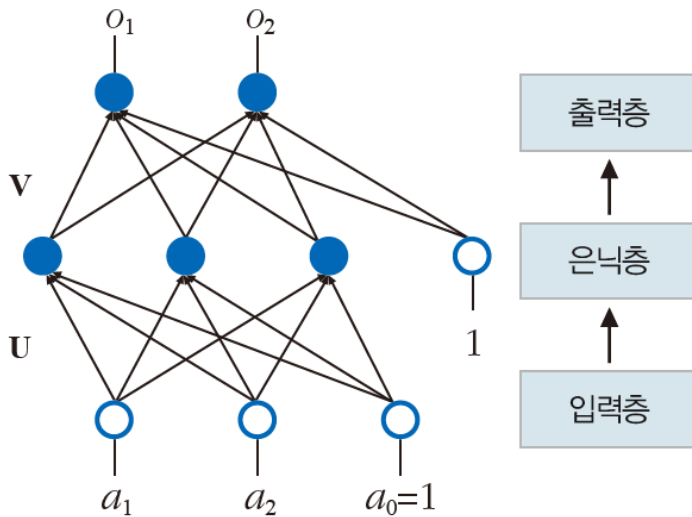
■ 시계열 데이터를 처리하는 신경망

- 시간에 따라 값이 하나씩 순차적으로 들어온다는 사실을 반영하는 신경망을 설계해야 함
- 다행히 다층 퍼셉트론을 약간 고쳐서 사용하면 됨

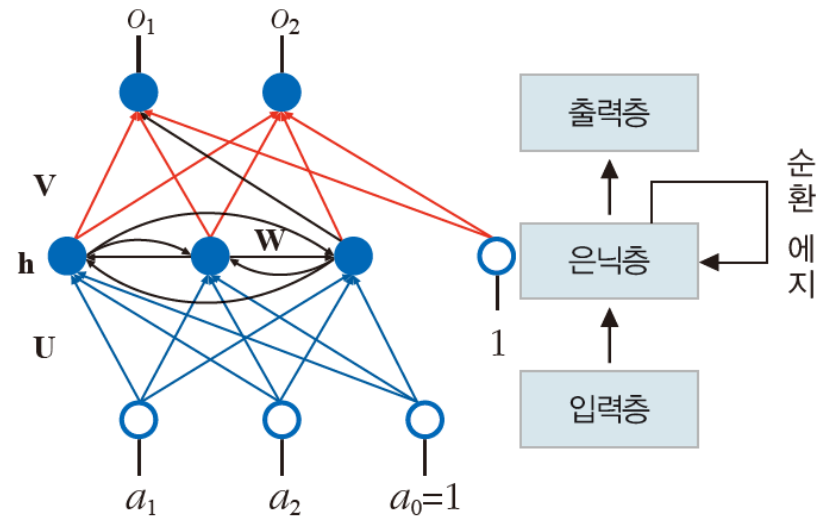
8.2.1 구조와 동작, 학습 알고리즘

■ 순환 신경망의 구조([그림 8-5(b)])

- 은닉층 노드 사이에 에지(순환 에지)가 있다는 사실을 제외하고 다층 퍼셉트론과 동일함



(a) 다층 퍼셉트론([그림 4-13])



(b) 순환 신경망([그림 4-13]에 순환 에지 추가)

그림 8-5 다층 퍼셉트론과 순환 신경망의 비교

- 가중치 집합
 - 다층 퍼셉트론 { \mathbf{U} , \mathbf{V} }
 - 순환 신경망 { \mathbf{U} , \mathbf{V} , \mathbf{W} }

8.2.1 구조와 동작, 학습 알고리즘

■ 순환 신경망에 데이터 입력([그림 8-6])

- 펼쳐서 그리면 이해가 쉬움
- i 순간에 \mathbf{a}^i 가 입력됨
- [그림 8-5(b)]의 a_0, a_1, a_2 가 \mathbf{a}^i 를 구성

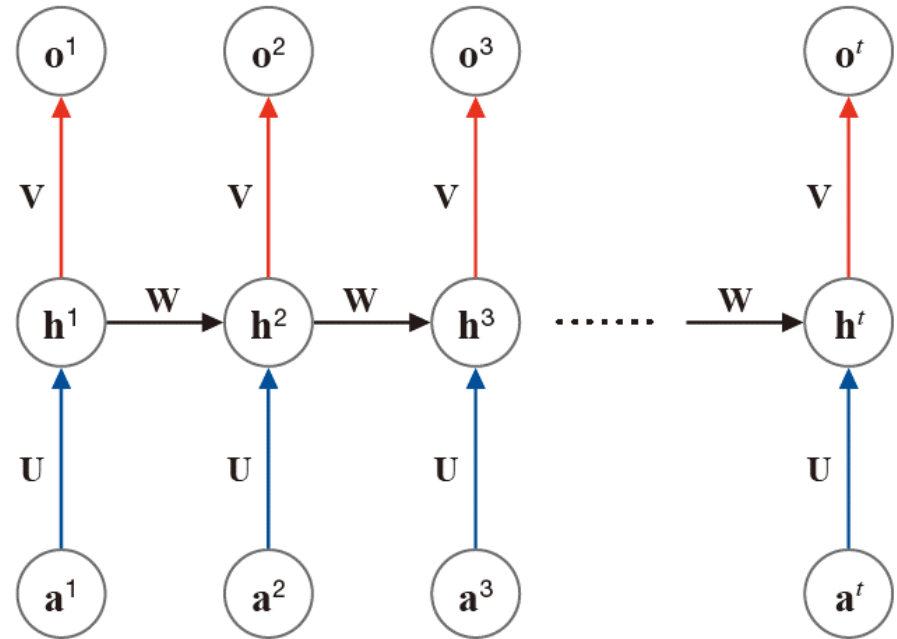


그림 8-6 펼쳐 순환 신경망

- 가중치 공유
 - 순간마다 서로 다른 가중치를 가지는 것이 아님
 - 모든 순간이 $\{\mathbf{U}, \mathbf{V}, \mathbf{W}\}$ 를 공유

8.2.1 구조와 동작, 학습 알고리즘

■ 순환 신경망의 동작

- i 순간의 \mathbf{a}^i 는 가중치 \mathbf{U} 를 통해 은닉층의 상태 \mathbf{h}^i 에 영향을 미치고, \mathbf{h}^i 는 가중치 \mathbf{V} 를 통해 출력값 \mathbf{o}^i 에 영향을 미침. \mathbf{h}^{i-1} 는 가중치 \mathbf{W} 를 통해 \mathbf{h}^i 에 영향을 미침

- 은닉층에서 일어나는 계산

$$\mathbf{h}^i = \tau_1(\mathbf{W}\mathbf{h}^{i-1} + \mathbf{U}\mathbf{a}^i) \quad (8.2)$$

- 출력층에서 일어나는 계산

$$\mathbf{o}^i = \tau_2(\mathbf{V}\mathbf{h}^i) \quad (8.3)$$

이 항을 제외하면 다층 퍼셉트론과 동일

- 식 (8.2)에서 $\mathbf{W}\mathbf{h}^{i-1}$ 항을 제외하면 다층 퍼셉트론과 동일

- 순환 신경망은 이 항을 통해 이전 순간의 은닉층 상태 \mathbf{h}^{i-1} 를 현재 순간의 은닉층 상태 \mathbf{h}^i 로 전달하여 시간성을 처리함

8.2.1 구조와 동작, 학습 알고리즘

■ 순환 신경망의 학습

- 학습 알고리즘은 최적의 $\{\mathbf{U}, \mathbf{V}, \mathbf{W}\}$ 를 알아냄
- BPTT(back-propagation through time) 알고리즘

8.2.2 선별 기억력을 갖춘 LSTM

■ 순환 신경망의 기억력 한계

- 은닉층 상태를 다음 순간으로 넘기는 기능을 통해 과거를 기억
- 하지만 장기 문맥 의존성(멀리 떨어진 요소가 밀접한 상호작용하는 현상)을 제대로 처리하지 못하는 한계
- 계속 들어오는 입력의 영향으로 기억력 감퇴
- 사람은 선별 기억 능력으로 오래 전 기억을 간직함

8.2.2 선별 기억력을 갖춘 LSTM

■ LSTM은 게이트라는 개념으로 선별 기억 확보

- o는 열림, x는 닫힘
- 실제로는 게이트는 0~1 사이의 실수값으로 열린 정도를 조절
- 게이트의 여닫는 정도는 가중치로 표현되며 가중치는 학습으로 알아냄

■ LSTM의 가중치

- 순환 신경망의 $\{U, V, W\}$ 에 4개를 추가하여 $\{U, U^i, U^o, W, W^i, W^o, V\}$
- i 는 입력 게이트, o 는 출력 게이트

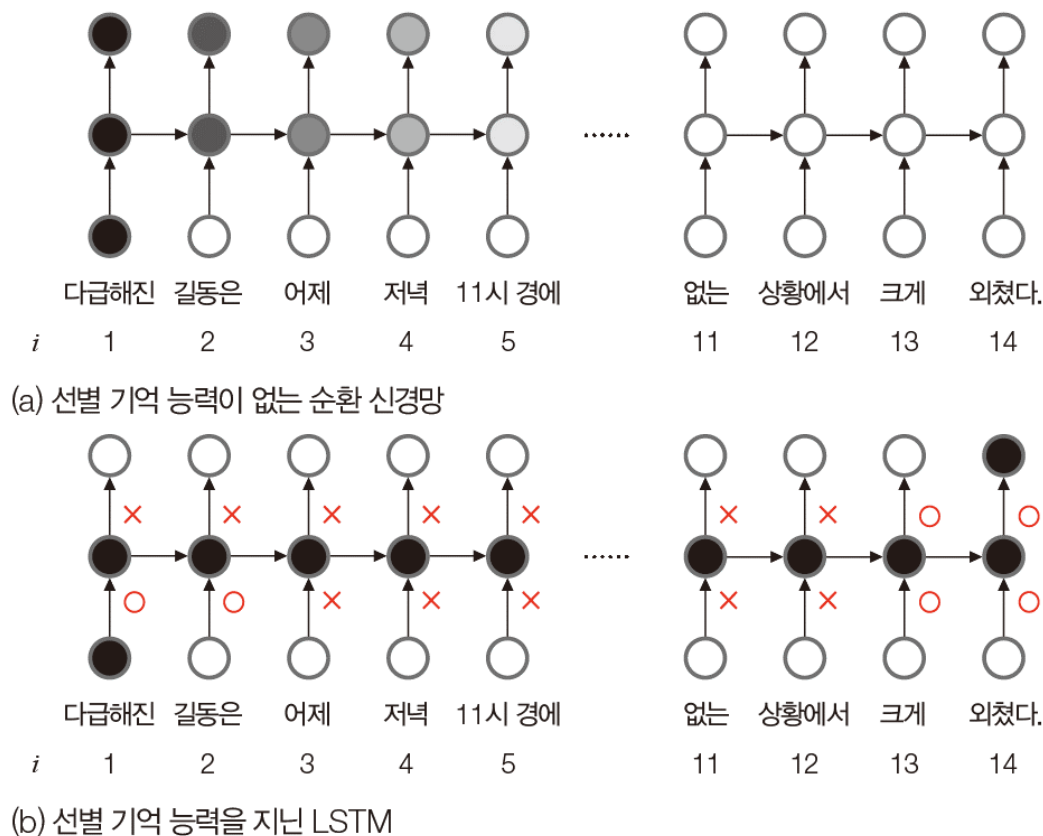
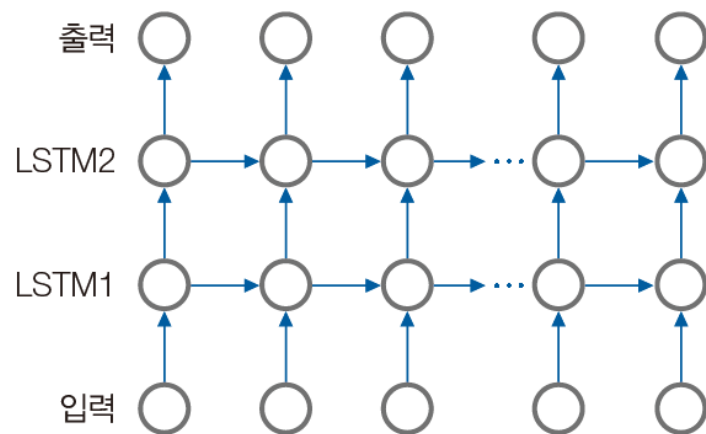


그림 8-7 순환 신경망과 LSTM

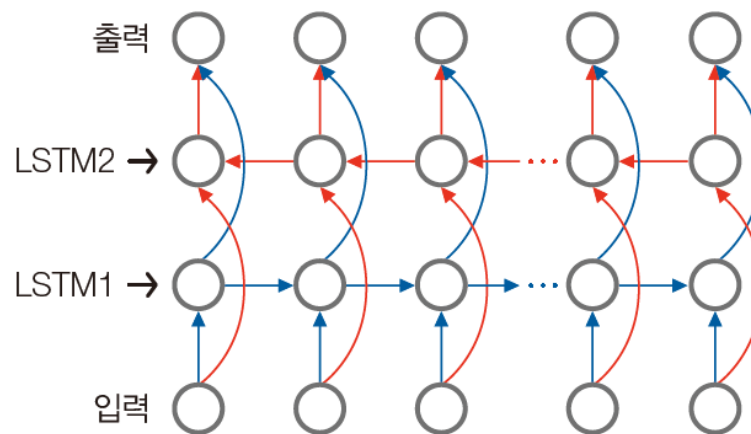
8.2.3 LSTM의 유연한 구조

■ 다양한 구조 설계 가능

- 양방향으로 문맥을 살필 필요가 있는 경우는 양방향 LSTM 사용
 - 예) "잘 달리는 이 차는 ..."과 "고산지대에서 생산한 이 차는 향기가 좋다"에서 앞 문장은 왼쪽, 뒤 문장은 오른쪽 단어를 보고 각각 car와 tea로 번역



(a) 적층 LSTM



(b) 양방향 LSTM

그림 8-8 적층 LSTM과 양방향 LSTM

8.3 LSTM으로 시계열 예측하기

- 시계열 데이터를 보고 미래를 예측하는 프로그래밍 실습
 - 단일 채널
 - 종가만 고려하는 [프로그램 8-2]
 - 다중 채널
 - 종가, 시가, 고가, 저가를 모두 고려하는 [프로그램 8-3]

8.3.1 단일 채널 비트코인 가격 예측

Train on 249 samples, validate on 108 samples

Epoch 1/200

249/249 - 3s - loss: 1858.8645 - mae: 1858.8646 - val_loss: 447.8866 - val_mae: 447.8866

...

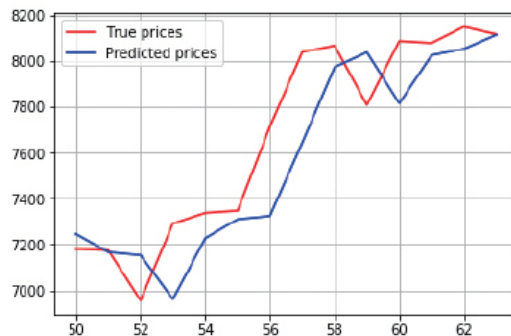
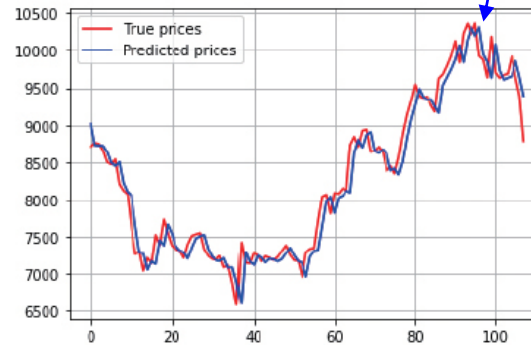
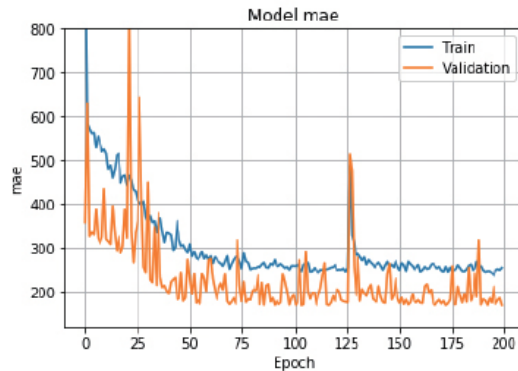
Epoch 200/200

249/249 - 1s - loss: 253.9756 - mae: 253.9756 - val_loss: 168.5357 - val_mae: 168.5357

손실 함수: 169.59739854600696 MAE: 169.5974
평균절댓값백분율오차(MAPE): [0.02028812]

MAE
MAPE

얼핏 아주 훌륭한 예측이라 보임



확대해 보면 파란 곡선(예측 값)은
빨간 곡선(정답)을 오른쪽으로 한 칸
이동한 모양.

8.3.1 단일 채널 비트코인 가격 예측

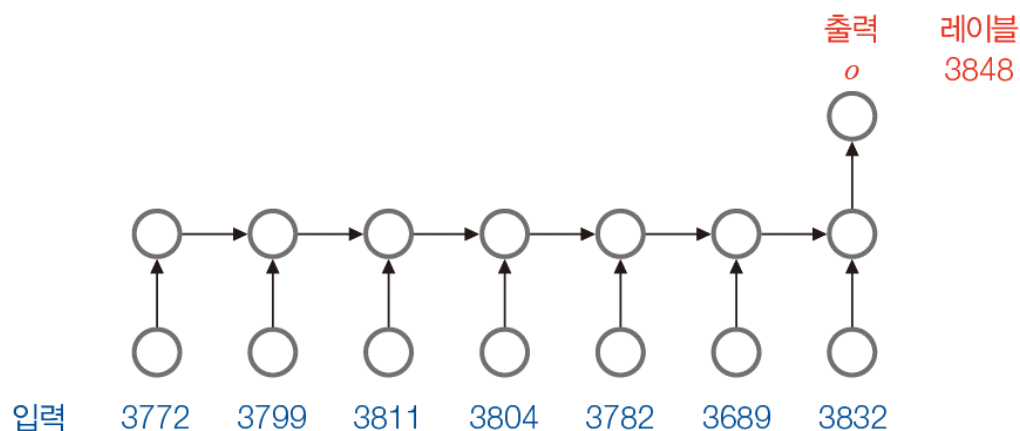


그림 8-10 [프로그램 8-2]의 신경망에 샘플 입력 후 출력과 레이블의 오차를 구하는 과정

NOTE 데이터 편향

[프로그램 8-2]의 실행 결과로 얻은 학습 곡선을 보면 검증 집합의 MAE가 훈련 집합보다 작은 기이한 현상을 볼 수 있다. [프로그램 8-1(a)]가 출력한 비트코인 가격의 원본 데이터를 보면, 초반부에 심한 등락이 나타나고 후반부는 상대적으로 안정된 추세를 보인다. 데이터를 7:3으로 나눌 때 앞부분 70%는 훈련 집합, 나머지 뒷부분은 테스트 집합에 담았다. 따라서 원천적으로 훈련 집합의 예측은 테스트 집합보다 어렵기 때문에 테스트 집합의 MAE가 더 작은 현상이 발생했다. 이처럼 데이터의 패턴이 한쪽으로 치우친 현상을 데이터 편향(data bias)이라고 한다. 의사와 간호사를 인식하는 인공지능 시스템을 학습했는데, 나중에 남자 간호사 사진을 입력하면 의사라고 분류하고 여자 의사 사진을 입력하면 간호사로 분류하는 현상도 데이터 편향 때문에 나타난다.

8.3.2 성능 평가

■ 성능 기준

- 평균절댓값오차 mean absolute error(MAE): 스케일 문제에 대처하지 못함

$$\text{평균절댓값오차: MAE} = \frac{1}{|M|} \sum_{x \in M} |y - o| \quad (8.4)$$

- 평균절댓값백분율오차: 스케일 문제에 대처

$$\text{평균절댓값백분율오차: MAPE} = \frac{1}{|M|} \sum_{x \in M} \left| \frac{y - o}{y} \right| \quad (8.5)$$

표 8-1 평균절댓값오차와 평균절댓값백분율오차

데이터 스케일	데이터(y_test는 참값, pred는 예측값)	평균절댓값오차(MAE)	평균절댓값백분율오차(MAPE)
두 자릿수	y_test [12 20 18 22 28]	(2+1+2+2+2)/5 =1.8	(2/12+1/20+2/18+2/22+2/28)/5 =0.0980
	pred [10 21 16 20 26]		
세 자릿수	y_test [120 200 180 220 280]	(20+10+20+20+20)/5 =18	(20/120+10/200+20/180+20/220+20/280)/5=0.0980
	pred [100 210 160 200 260]		

8.3.2 성능 평가

■ 성능 기준

■ 등락 정확률

- 등락을 얼마나 정확하게 맞히는지 측정
- 맞힌 경우의 수를 전체 샘플 수로 나눔

표 8-2 등락 정확률

샘플 i	x_test[i]	y_test[i]	pred[i]	맞힘
1	[.,, .., .., 21]	23	24	o
2	[.,, .., .., 25]	20	26	x
3	[.,, .., .., 22]	24	23	o
4	[.,, .., .., 28]	25	26	o
5	[.,, .., .., 21]	18	17	o
6	[.,, .., .., 32]	31	33	x
7	[.,, .., .., 35]	36	37	o
8	[.,, .., .., 20]	19	22	x
				등락 정확률 = 5/8 = 62.5%

8.3.3 다중 채널 비트코인 가격 예측

훈련 집합 X와 Y의 텐서 모양

(358, 7, 4) (358, 4)

```
[[3772.93633533 3796.63728431 3824.16587937 3666.52401643]
 [3799.67854295 3773.44146075 3879.23118467 3753.80002246]
 [3811.61197937 3799.36702601 3840.04482307 3788.91849833]
 [3804.41917011 3806.69151279 3819.19435612 3759.40921647]
 [3782.66410112 3807.84575592 3818.69548135 3766.24204823]
 [3689.86289319 3783.35506344 3804.35361623 3663.47774336]
 [3832.08088473 3701.04987103 3866.71870424 3688.69715385]]
 [3848.95636968 3832.59242908 3881.96576977 3802.51605364]
```

Train on 249 samples, validate on 108 samples

Epoch 1/200

249/249 - 3s - loss: 1527.8208 - mae: 1527.8209 - val_loss: 329.5129 - val_mae: 329.5129

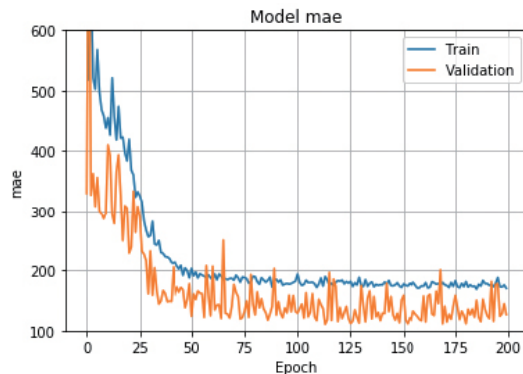
...

Epoch 200/200

249/249 - 1s - loss: 170.3932 - mae: 170.3933 - val_loss: 126.5666 - val_mae: 126.5666

손실 함수: 126.56658257378473 MAE: 126.56658

LSTM 평균절댓값백분율오차(MAPE): [0.02284203 0.00305788 0.01866189 0.01706315]



(종가,시가,고가,저가)의 MAPE

8.5 자연어 처리

■ 언어를 사용하는 인간

- 다른 동물보다 지능이 월등하다는 증거
- 시를 읽고 반응하는 인간

연탄재 함부로 발로 차지 마라

너는

누구에게 한 번이라도 뜨거운 사람이었느냐

■ 자연어 처리_{natural language processing(NLP)}

- 인간이 구사하는 언어를 자동으로 처리하는 인공지능 분야
- 다양한 응용
 - 언어 번역
 - 영화평 댓글 분석하여 흥행 추정
 - 고객 응대 챗봇
 - 소설이나 시를 쓰는 창작 인공지능 등

8.5.1 텍스트 데이터에 대한 이해

■ 영화평 데이터셋인 IMDB의 예제 문장

- 텍스트의 특성이 잘 나타남

Once again Mr. Costner has dragged out a movie for far longer than necessary. Aside from the terrific sea rescue sequences, of which there are very few I just did not care about any of the characters. Most of us have ghosts in the closet, and Costner's character are realized early on, and then forgotten until much later, by which time I did not care,

■ 텍스트 데이터의 특성

- 시계열 데이터로서 시간 정보가 있고 샘플마다 길이가 다르다는 기본 성질
- 그 외 독특한 특성
 - 심한 잡음
 - 형태소 분석 필요
 - 구문론과 의미론
 - 다양한 언어 특성
 - 신경망에 입력하려면 기호를 수치로 변환해야 함

8.5.1 텍스트 데이터에 대한 이해

■ 원핫 코드 표현으로 변환하는 절차

- 예) 말뭉치_{corpus} 사례

[말뭉치]

```
Freshman loves python.  
We teach python to freshman.  
How popular is Python?
```

- 단어 수집(괄호 속은 빈도수)
 - Python(3), freshman(2), loves(1), we(1), teach(1), to(1), how(1), popular(1), is(1)
- 파이썬의 자료구조인 딕셔너리를 이용한 표현(빈도수에 따른 순위 부여)
 - {'python':1, 'freshman':2, 'loves':1, 'we':1, 'teach':1, 'to':1, 'how':1, 'popular':1, 'is':1}
- 텍스트를 숫자 코드로 변환

Freshman loves python.	→	[2 3 1]
We teach python to freshman.	→	[4 5 1 6 2]
How popular is Python?	→	[7 8 9 1]

8.5.1 텍스트 데이터에 대한 이해

■ 원핫 코드 표현

```
[231]    → [[010000000] [001000000] [100000000]]  
[45162] → [000100000] [000010000] [100000000] [000001000] [010000000]  
[7891]  → [000000100] [000000010] [000000001] [100000000]]
```

■ 원핫 코드의 문제점과 해결책

- 사전 크기가 크면 원핫 코드는 희소 벡터가 되어 메모리 낭비
- 단어 사이의 연관 관계를 반영하지 못함
 - 예, king과 queen, bridegroom과 bride의 관계 표현 불가능

8.5.2 텍스트 데이터 사례: 영화평 데이터셋 IMDB

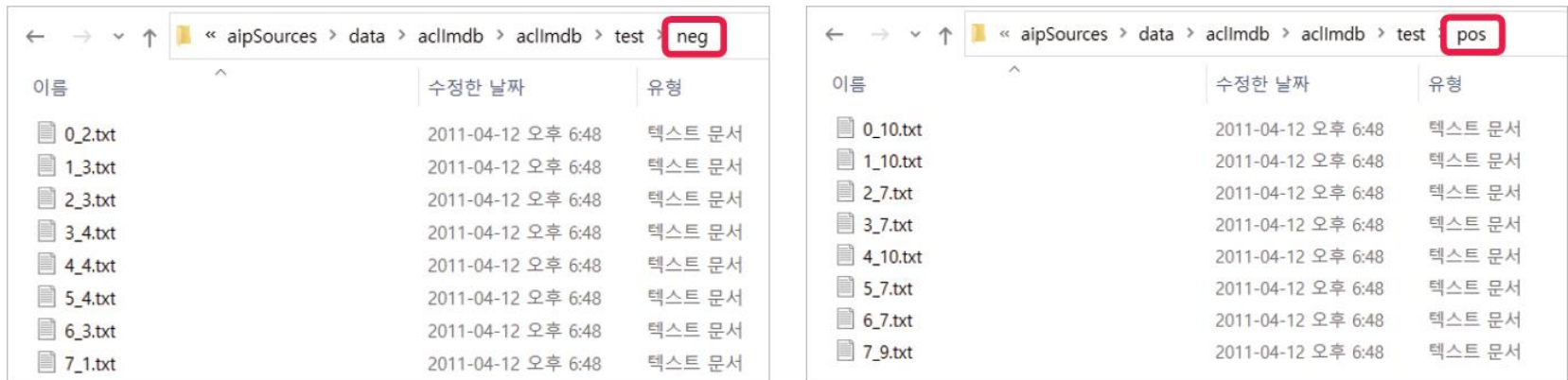
■ 텐서플로가 제공하는 텍스트 데이터

- 영화를 평가한 댓글을 모아둔 IMDB 데이터셋
 - 50000개의 댓글을 긍정 평가와 부정 평가로 레이블링
 - 감정 분류 sentiment classification 문제에 주로 사용
- 다양한 토픽의 뉴스를 모아둔 Reuters 데이터셋
 - 로이터 통신의 뉴스 11228개를 46개 토픽으로 레이블링
 - 토픽 분류 문제에 주로 사용

8.5.2 텍스트 데이터 사례: 영화평 데이터셋 IMDB

■ IMDB로 실습

- <http://mng.bz/0tlo>에 접속하여 원본 데이터 다운로드
- 소스 프로그램이 있는 폴더에 data라는 폴더 만들어 거기에 저장



이름	수정한 날짜	유형
0_2.txt	2011-04-12 오후 6:48	텍스트 문서
1_3.txt	2011-04-12 오후 6:48	텍스트 문서
2_3.txt	2011-04-12 오후 6:48	텍스트 문서
3_4.txt	2011-04-12 오후 6:48	텍스트 문서
4_4.txt	2011-04-12 오후 6:48	텍스트 문서
5_4.txt	2011-04-12 오후 6:48	텍스트 문서
6_3.txt	2011-04-12 오후 6:48	텍스트 문서
7_1.txt	2011-04-12 오후 6:48	텍스트 문서

이름	수정한 날짜	유형
0_10.txt	2011-04-12 오후 6:48	텍스트 문서
1_10.txt	2011-04-12 오후 6:48	텍스트 문서
2_7.txt	2011-04-12 오후 6:48	텍스트 문서
3_7.txt	2011-04-12 오후 6:48	텍스트 문서
4_10.txt	2011-04-12 오후 6:48	텍스트 문서
5_7.txt	2011-04-12 오후 6:48	텍스트 문서
6_7.txt	2011-04-12 오후 6:48	텍스트 문서
7_9.txt	2011-04-12 오후 6:48	텍스트 문서

그림 8-16 IMDB 데이터셋의 폴더 구조 - neg와 pos 폴더

8.5.2 텍스트 데이터 사례: 영화평 데이터셋 IMDB

훈련 집합과 테스트 집합의 크기

(25000,) (25000,)

훈련 집합의 첫 번째 샘플

[1, 14, 22, 16, 43, 530, 973, 1622, 1385, 65, 458, 4468, 66, 3941, 4, 173, 36, 256, 5, 25, 100, 43, 838, 112, 50, 670, 2, 9, 35, 480, 284, 5, 150, 4, 172, 112, 167, 2, 336, 385, 39, 4, 172, 4536, 1111, 17, 546, 38, 13, 447, 4, 192, 50, 16, 6, 147, 2025, 19, 14, 22, 4, 1920, 4613, 469, 4, 22, 71, 87, 12, 16, 43, 530, 38, 76, 15, 13, 1247, 4, 22, 17, 515, 17, 12, 16, 626, 18, 2, 5, 62, 386, 12, 8, 316, 8, 106, 5, 4, 2223, 5244, 16, 480, 66, 3785, 33, 4, 130, 12, 16, 38, 619, 5, 25, 124, 51, 36, 135, 48, 25, 1415, 33, 6, 22, 12, 215, 28, 77, 52, 5, 14, 407, 16, 82, 2, 8, 4, 107, 117, 5952, 15, 256, 4, 2, 7, 3766, 5, 723, 36, 71, 43, 530, 476, 26, 400, 317, 46, 7, 4, 2, 1029, 13, 104, 88, 4, 381, 15, 297, 98, 32, 2071, 56, 26, 141, 6, 194, 7486, 18, 4, 226, 22, 21, 134, 476, 26, 480, 5, 144, 30, 5535, 18, 51, 36, 28, 224, 92, 25, 104, 4, 226, 65, 16, 38, 1334, 88, 12, 16, 283, 5, 16, 4472, 113, 103, 32, 15, 16, 5345, 19, 178, 32]

the/and/a/of/to/is/br/in/it/i/this/that/was/as/for/with/movie/but/film/on/

상위 빈도 20개 단어

8.5.3 단어 임베딩

■ 단어 임베딩이란?

- 단어를 저차원 공간의 벡터로 표현하는 기법
 - 보통 수백 차원을 사용
 - 밀집 벡터임
 - 단어의 의미를 표현
 - 신경망 학습을 통해 알아냄

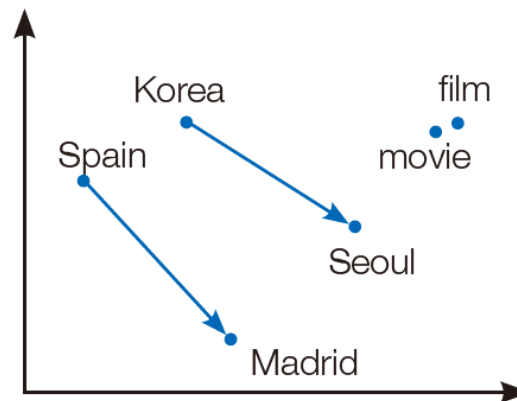


그림 8-17 단어 임베딩의 원리

[원핫 코드]

사전 크기([프로그램 8-7(a)]의 `dic_siz`)가 10000인 경우

"python" → $\overbrace{(1\ 0\ 0\ 0\ 0\ \dots\ 0\ 0\ 0\ \dots\ 0\ 0\ 0)}^{10000\text{개 요소}}$

[단어 임베딩]

단어 임베딩 공간의 차원([프로그램 8-7(b)]의 `embed_space_dim`)이 16인 경우

"python" → $\overbrace{(0.01\ 0.23\ 0.0\ \dots\ 0.002)}^{16\text{개 요소}}$

8.5.4 LSTM으로 인식: 조기 멈춤 적용

■ LSTM을 사용하는 프로그램

- 시계열 정보를 반영(단지 단어 임베딩에 다층 퍼셉트론을 적용하여 시계열 정보를 이용하지 못함)
- 조기 멈춤 적용(실행 결과를 보면 세대 1에서 최고 성능을 이룬 후 개선 없음). 조기 멈춤 (Early Stopping)은 훈련 집합에 대해 덜 수렴했더라도 검증 집합에 대해 성능 개선이 없으면 학습을 마치는 전략

8.5.4 LSTM으로 인식: 조기 멈춤 적용

Train on 20000 samples, validate on 5000 samples

Epoch 1/20

20000/20000 - 119s - loss: 0.4790 - accuracy: 0.7628 - val_loss: 0.3398 - val_accuracy: 0.8640

Epoch 2/20

20000/20000 - 113s - loss: 0.2587 - accuracy: 0.9004 - val_loss: 0.3117 - val_accuracy: 0.8722

Epoch 3/20

20000/20000 - 109s - loss: 0.1976 - accuracy: 0.9288 - val_loss: 0.3347 - val_accuracy: 0.8784

Epoch 4/20

20000/20000 - 119s - loss: 0.1684 - accuracy: 0.9401 - val_loss: 0.3417 - val_accuracy: 0.8754

Epoch 5/20

20000/20000 - 107s - loss: 0.1239 - accuracy: 0.9588 - val_loss: 0.3424 - val_accuracy: 0.8682

Epoch 6/20

20000/20000 - 105s - loss: 0.0987 - accuracy: 0.9685 - val_loss: 0.4095 - val_accuracy: 0.8734

Epoch 7/20

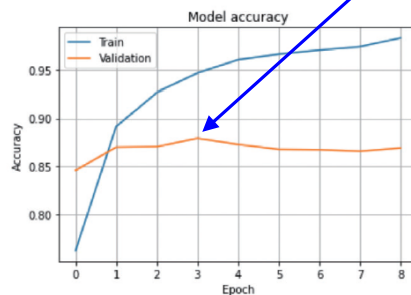
20000/20000 - 105s - loss: 0.0908 - accuracy: 0.9709 - val_loss: 0.4010 - val_accuracy: 0.8618

Epoch 8/20

20000/20000 - 104s - loss: 0.1116 - accuracy: 0.9596 - val_loss: 0.4849 - val_accuracy: 0.8638

정확률은 86.76400184631348

세대 3 이후 다섯 세대 동안 성능 향상이 없어 학습을 멈춤



8.5.4 LSTM으로 인식: 조기 멈춤 적용

■ 성능 분석

- LSTM을 사용해 시계열 특성을 반영하고 조기 멈춤을 적용해 더 유리한 상황인데 이전 프로그램 보다 열등. 왜?
 - 두 프로그램 모두 단어의 빈도수에 따라 분류하는 듯
 - boring, terrible, bad 등이 많으면 부정, wonderful, good 등이 자주 나타나면 긍정으로 분류
 - 문장의 의미를 파악하지 못한 채 분류
 - 예, "To me all of the movies are terrible, but this one is not." 문장에서 terrible이 있다는 이유로 두 프로그램 모두 부정으로 분류

■ 문장의 의미 이해하려면 발전된 자연어 처리 알고리즘 필요

- word2vec과 GloVe

8.5.5 word2vec과 GloVe

■ 단어를 벡터 공간에 표현하는 단어 임베딩 기술

- 오래 전부터 연구되어온 아이디어
- 영국의 언어학자 퍼스 "You shall know a word by the company it keeps." 단어 간의 상호 작용이 매우 중요하다는 통찰
 - 예) "영화"라는 단어는 "아카데미", "흥행" 등의 단어와 함께 등장할 가능성 높음
- 고전적인 기법들: TF(term frequency), LSA, 신경망 기법들
- 2010년대에는 딥러닝을 활용한 단어 임베딩 기술이 주류
 - Word2vec(구글)
 - 1000억개 가량의 뉴스를 모아둔 데이터셋으로 학습. 300만개 가량의 단어를 300차원 공간에 표현
 - GloVe(스탠퍼드 대학)
 - 위키피디아 문서 데이터를 사용하여 학습. 40만개 가량의 단어를 50, 100, 200, 300 차원 공간에 표현

8.5.5 word2vec과 GloVe

- GloVe로 단어 연관 관계를 찾는 프로그래밍 실습
 - 파일 크기가 상대적으로 작은 GloVE를 가지고 실습
 - 100차원 파일인 glove.6B.100d.txt 사용

8.5.5 word2vec과 GloVe

