2022학년도 1학기 컴퓨터언어학

제16강 시퀀스 처리 (3)

박수지

서울대학교 인문대학 언어학과

2022년 5월 2일 월요일

오늘의 목표

1. 단순 순환 신경망의 역전파에서 일어나는 기울기 소실 문제의 두 가지 원인을 설명할 수 있다.
2. 장-단기 기억 순환 신경망(LSTM)의 구조를 이해하고 각 게이트의 역할을 설명할 수 있다.

순환 신경망의 계산 그래프

순전파(길이 3인 경우)

순환 신경망의 은닉층

$$\vec{h}_t = \tanh\left(\mathbf{U}\vec{h}_{t-1} + \mathbf{W}\vec{x}_t\right)$$

$$\vec{a}_1 = \mathbf{U}\vec{h}_0 \qquad \vec{a}_2 = \mathbf{U}\vec{h}_1 \qquad \vec{a}_3 = \mathbf{U}\vec{h}_2$$

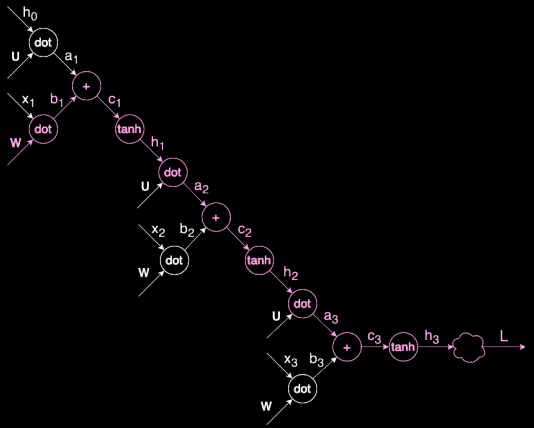
$$\vec{b}_1 = \mathbf{W}\vec{x}_1 \qquad \vec{b}_2 = \mathbf{W}\vec{x}_2 \qquad \vec{b}_3 = \mathbf{W}\vec{x}_3$$

$$\vec{c}_1 = \vec{a}_1 + \vec{b}_1 \qquad \vec{c}_2 = \vec{a}_2 + \vec{b}_2 \qquad \vec{c}_3 = \vec{a}_3 + \vec{b}_3$$

$$\vec{h}_1 = \tanh \vec{c}_1 \qquad \vec{h}_2 = \tanh \vec{c}_2 \qquad \vec{h}_3 = \tanh \vec{c}_3$$

순환 신경망의 계산 그래프

역전파(길이 3인 경우)



연쇄 법칙

$$\frac{\partial \mathrm{L}}{\partial \mathbf{W}} = \frac{\partial \mathrm{L}}{\partial \vec{\mathrm{h}}_3} \cdot \frac{\partial \vec{\mathrm{h}}_3}{\partial \vec{\mathrm{c}}_3} \cdot \frac{\partial \vec{\mathrm{c}}_3}{\partial \vec{\mathrm{a}}_3} \cdot \frac{\partial \vec{\mathrm{a}}_3}{\partial \vec{\mathrm{h}}_2} \cdot \frac{\partial \vec{\mathrm{h}}_2}{\partial \vec{\mathrm{c}}_2} \cdot \frac{\partial \vec{\mathrm{c}}_2}{\partial \vec{\mathrm{a}}_2} \cdot \frac{\partial \vec{\mathrm{a}}_2}{\partial \vec{\mathrm{h}}_1} \cdot \frac{\partial \vec{\mathrm{h}}_1}{\partial \vec{\mathrm{c}}_1} \cdot \frac{\partial \vec{\mathrm{c}}_1}{\partial \vec{\mathrm{b}}_1} \cdot \frac{\partial \vec{\mathrm{b}}_1}{\partial \mathbf{W}}$$

순환 신경망의 계산 그래프

역전파(길이 3인 경우)

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial \vec{h}_3} \cdot \frac{\partial \vec{h}_3}{\partial \vec{c}_3} \cdot \frac{\partial \vec{c}_3}{\partial \vec{a}_3} \cdot \frac{\partial \vec{a}_3}{\partial \vec{h}_2} \cdot \frac{\partial \vec{h}_2}{\partial \vec{c}_2} \cdot \frac{\partial \vec{c}_2}{\partial \vec{a}_2} \cdot \frac{\partial \vec{a}_2}{\partial \vec{h}_1} \cdot \frac{\partial \vec{h}_1}{\partial \vec{c}_1} \cdot \frac{\partial \vec{c}_1}{\partial \vec{b}_1} \cdot \frac{\partial \vec{b}_1}{\partial \mathbf{W}}$$

$$\frac{\partial L}{\partial \vec{h}_3} = \begin{bmatrix} \frac{\partial L}{\partial h_3^1} \\ \vdots \\ \frac{\partial L}{\partial h_3^{d_h}} \end{bmatrix} \in \mathbb{R}^{d_h}$$

순환 신경망의 계산 그래프

역전파(길이 3인 경우)

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial \vec{h}_3} \cdot \frac{\partial \vec{h}_3}{\partial \vec{c}_3} \cdot \frac{\partial \vec{c}_3}{\partial \vec{a}_3} \cdot \frac{\partial \vec{a}_3}{\partial \vec{h}_2} \cdot \frac{\partial \vec{h}_2}{\partial \vec{c}_2} \cdot \frac{\partial \vec{c}_2}{\partial \vec{a}_2} \cdot \frac{\partial \vec{a}_2}{\partial \vec{h}_1} \cdot \frac{\partial \vec{h}_1}{\partial \vec{c}_1} \cdot \frac{\partial \vec{c}_1}{\partial \vec{b}_1} \cdot \frac{\partial \vec{b}_1}{\partial \mathbf{W}}$$

$$\vec{h}_t = \tanh \vec{c}_t \quad \Rightarrow \quad \frac{\partial \vec{h}_t}{\partial \vec{c}_t} = \vec{1} - \tanh^2 \vec{c}_t \in \mathbb{R}^{d_h}$$

순환 신경망의 계산 그래프

역전파(길이 3인 경우)

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial \vec{h}_3} \cdot \frac{\partial \vec{h}_3}{\partial \vec{c}_3} \cdot \frac{\partial \vec{c}_3}{\partial \vec{a}_3} \cdot \frac{\partial \vec{a}_3}{\partial \vec{h}_2} \cdot \frac{\partial \vec{h}_2}{\partial \vec{c}_2} \cdot \frac{\partial \vec{c}_2}{\partial \vec{a}_2} \cdot \frac{\partial \vec{a}_2}{\partial \vec{h}_1} \cdot \frac{\partial \vec{h}_1}{\partial \vec{c}_1} \cdot \frac{\partial \vec{c}_1}{\partial \vec{b}_1} \cdot \frac{\partial \vec{b}_1}{\partial \mathbf{W}}$$

$$\vec{c}_t = \vec{a}_t + \vec{b}_t \quad \Rightarrow \quad \frac{\partial \vec{c}_t}{\partial \vec{a}_t} = \vec{1}, \frac{\partial \vec{c}_t}{\partial \vec{b}_t} = \vec{1} \in \mathbb{R}^{d_h}$$

순환 신경망의 계산 그래프

역전파(길이 3인 경우)

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial \vec{h}_3} \cdot \frac{\partial \vec{h}_3}{\partial \vec{c}_3} \cdot \frac{\partial \vec{c}_3}{\partial \vec{a}_3} \cdot \frac{\partial \vec{a}_3}{\partial \vec{h}_2} \cdot \frac{\partial \vec{h}_2}{\partial \vec{c}_2} \cdot \frac{\partial \vec{c}_2}{\partial \vec{a}_2} \cdot \frac{\partial \vec{a}_2}{\partial \vec{h}_1} \cdot \frac{\partial \vec{h}_1}{\partial \vec{c}_1} \cdot \frac{\partial \vec{c}_1}{\partial \vec{b}_1} \cdot \frac{\partial \vec{b}_1}{\partial \mathbf{W}}$$

$$\vec{a}_t = \mathbf{U}\vec{h}_{t-1} \quad \Rightarrow \quad \frac{\partial \vec{a}_t}{\partial \vec{h}_{t-1}} = \mathbf{U} \in \mathbb{R}^{d_h \times d_h}$$

이유: 예시

$$\vec{y} = \mathbf{A}\vec{x} \;\Rightarrow\; \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \;\Rightarrow\; \frac{\partial \vec{y}}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \\ \frac{\partial y_3}{\partial x_1} & \frac{\partial y_3}{\partial x_2} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} = \mathbf{A}$$

순환 신경망의 계산 그래프

역전파(길이 3인 경우)

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial \vec{h}_3} \cdot \frac{\partial \vec{h}_3}{\partial \vec{c}_3} \cdot \frac{\partial \vec{c}_3}{\partial \vec{a}_3} \cdot \frac{\partial \vec{a}_3}{\partial \vec{h}_2} \cdot \frac{\partial \vec{h}_2}{\partial \vec{c}_2} \cdot \frac{\partial \vec{c}_2}{\partial \vec{a}_2} \cdot \frac{\partial \vec{a}_2}{\partial \vec{h}_1} \cdot \frac{\partial \vec{h}_1}{\partial \vec{c}_1} \cdot \frac{\partial \vec{c}_1}{\partial \vec{b}_1} \cdot \frac{\partial \vec{b}_1}{\partial \mathbf{W}}$$

$$\vec{b}_t = \mathbf{W}\vec{x}_t \quad \Rightarrow \quad \frac{\partial \vec{b}_t}{\partial \mathbf{W}} \in \mathbb{R}^{d_h \times d_{in}}$$

순환 신경망의 계산 그래프

역전파(길이 3인 경우)

관찰

연쇄의 첫 단계에서 가중치 매개변수 $\mathbf{W}$에 대한 손실함수 L의 편도함수 $\frac{\partial L}{\partial \mathbf{W}}$를 구할 때

1. $\frac{\partial \vec{h}_t}{\partial \vec{c}_t} = \left(\vec{1} - \tanh^2 \vec{c}_t\right)$ 벡터의 성분별 곱셈이 3회 이루어진다.
2. $\frac{\partial \vec{a}_t}{\partial \vec{h}_{t-1}} = \mathbf{U}$의 행렬곱이 2회 이루어진다.

순환 신경망의 계산 그래프

역전파

일반화

길이 n인 연쇄의 첫 단계에서 $\frac{\partial L}{\partial \mathbf{W}}$를 구할 때

1. $\frac{\partial \vec{h}_t}{\partial \vec{c}_t} = \left(\vec{1} - \tanh^2 \vec{c}_t\right)$ 벡터의 성분별 곱셈이 n회 이루어진다.
2. $\frac{\partial \vec{a}_t}{\partial \vec{h}_{t-1}} = \mathbf{U}$의 행렬곱이 $(n-1)$회 이루어진다.

기울기 소실(Vanishing gradient)

n이 커질 때 일어나는 일

1. 모든 실수 c에 대해 $\left|1 - \tanh^2 c\right| \leq 1$이므로 많이 곱할수록 크기가 0에 가까워진다.
2. $|\mathbf{U}| < 1$일 때 제곱을 반복하면 크기가 0에 아주 가까워지고
   $|\mathbf{U}| > 1$일 때 제곱을 반복하면 크기가 무한대로 발산한다.

계산상의 문제: $n = 162$만 되어도...

```
>>> 0.01 ** 162
0.0
```

문제점

x_1의 정보가 연쇄 끝까지 전달되지 못하고 0이 되어 사라진다.

단순 순환 신경망의 문제점

1. 은닉층에서 두 가지 일을 동시에 수행해야 한다.
   1. 현재 상태의 결정에 필요한 정보 $\vec{y}_t = \text{softmax}\left(\mathbf{V}\vec{h_t}\right)$
   2. 미래 상태의 결정에 필요한 정보 $\vec{h}_{t+1} = \tanh\left(\mathbf{U}\vec{h_t} + \mathbf{W}\vec{x_{t+1}}\right)$
2. 기울기 소실 또는 폭발이 일어난다.
   - $\tanh$ 함수의 기울기
   - $\mathbf{U}$의 반복된 곱셈

기울기 소실 문제의 해결

정보를 저장해서 나중에 사용한다. $\Rightarrow$ Long-Short Term Memory

장·단기 기억(LSTM)의 아이디어

맥락을 다루는 문제를 두 개로 분리한다.

1. 불필요한 정보를 삭제하기 $\Rightarrow$ Forget gate
2. 이후의 결정에 필요한 정보를 더하기 $\Rightarrow$ Add(Input) gate

LSTM의 새로운 요소

1. 명시적인 문맥 계층 $\vec{c}_t$
2. 정보의 흐름을 통제하는 게이트

LSTM 단위

- 입력
  - $\vec{c}_{t-1}$ 직전의 문맥 상태(context/carry state)
  - $\vec{h}_{t-1}$ 직전의 은닉 상태(hidden state)
  - $\vec{x}_t$ 현재의 입력
- 출력
  - $\vec{c}_t$ 현재의 문맥 상태 (미래의 결정을 위해 저장하는 정보)
  - $\vec{h}_t$ 현재의 은닉 상태 (현재의 결정에 사용하는 정보)

LSTM 게이트의 종류

삭제(Forget) 과거의 문맥 상태($\vec{c}_{t-1}$) 중에서 삭제할 정보를 선택한다.

추가(Add/Input) $\vec{h}_{t-1}$과 $\vec{x}_t$로부터 현재의 문맥 상태($\vec{c}_t$)에 추가할 정보를 선택한다.

출력(Output) $\vec{h}_{t-1}$과 $\vec{x}_t$로부터 현재의 은닉 상태($\vec{h}_t$)에 필요한 정보를 결정한다.

게이트의 목적

정보의 취사 선택: [게이트를 통과한 정보] $=$ [실제 정보] $\odot$ [마스크]

"마스크" 모든 성분이 0과 1 사이의 실수 값으로 이루어진 벡터.
0이면 정보를 완전히 버리고, 1이면 정보를 온전히 가져간다.

$\odot$ 배열의 성분별 곱셈

게이트의 작용

[게이트를 통과한 정보] = [실제 정보] $\odot$ [마스크]

- $\vec{c}_t = \vec{c}_{t-1} \odot \vec{f}_t + \vec{g}_t \odot \vec{i}_t$ ⋯ 미래의 결정을 위해 저장하는 정보
  - c_{t-1} 과거의 문맥 상태에 관한 정보 (삭제 게이트 통과)
  - f_t 삭제 게이트의 마스크 ⋯ 과거의 정보를 얼마나 삭제할 것인가?
  - g_t 현재의 문맥 상태에 추가할 정보 (추가/입력 게이트 통과)
  - i_t 추가/입력 게이트의 마스크 ⋯ 현재의 정보를 얼마나 추가할 것인가?
- $\vec{h}_t = \tanh \vec{c}_t \odot \vec{o}_t$ ⋯ 현재의 결정에 사용하는 정보
  - $\tanh \vec{c}_t$ 현재의 은닉 상태에 추가할 정보 (출력 게이트 통과)
  - $\vec{o}_t$ 출력 게이트의 마스크

$$\vec{c}_t = \vec{c}_{t-1} \odot \vec{f}_t + \vec{g}_t \odot \vec{i}_t$$

$$\vec{h}_t = \tanh \vec{c}_t \odot \vec{o}_t$$

삭제 게이트

정보 $\vec{c}_{t-1}$

마스크 $\vec{f}_t = \sigma\left(\mathbf{U}_f \vec{h}_{t-1} + \mathbf{W}_f \vec{x}_t\right)$

추가/입력 게이트

정보 $\vec{g}_t = \tanh\left(\mathbf{U}_g \vec{h}_{t-1} + \mathbf{W}_g \vec{x}_t\right)$

마스크 $\vec{i}_t = \sigma\left(\mathbf{U}_i \vec{h}_{t-1} + \mathbf{W}_i \vec{x}_t\right)$

출력 게이트

정보 $\tanh \vec{c}_t$

마스크 $\vec{o}_t = \sigma\left(\mathbf{U}_o \vec{h}_{t-1} + \mathbf{W}_o \vec{x}_t\right)$

$$\vec{x}_t \in \mathbb{R}^{d_{in}}$$

$$\vec{c}_t, \vec{h}_t \in \mathbb{R}^{d_h}$$

$$\mathbf{U}_f, \mathbf{U}_g, \mathbf{U}_i, \mathbf{U}_o \in \mathbb{R}^{d_h \times d_h}$$

$$\mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_i, \mathbf{W}_o \in \mathbb{R}^{d_h \times d_{in}}$$

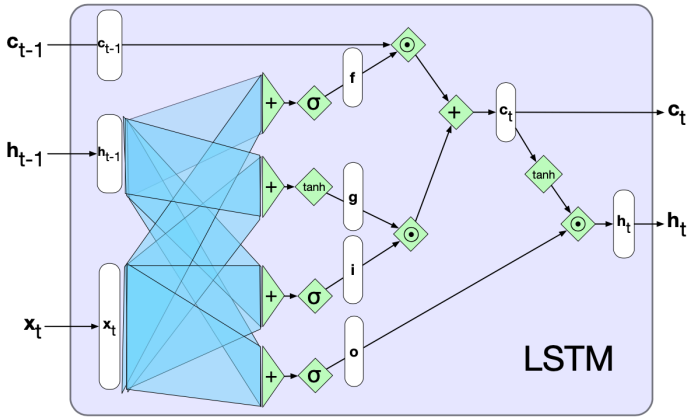
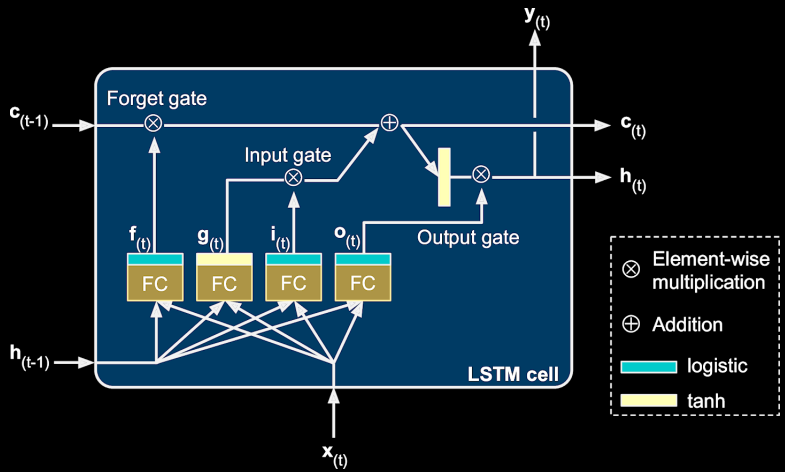


Figure 9.13 A single LSTM unit displayed as a computation graph. The inputs to each unit consists of the current input, x, the previous hidden state, h_{t-1}, and the previous context, c_{t-1}. The outputs are a new hidden state, h_t and an updated context, c_t.



Aurélien Géron. Neural networks and deep learning. O'Reilly. 2018.

https://www.oreilly.com/library/view/neural-networks-and/9781492037354/assets/mlst_1413.png

순환 신경망의 근본적인 한계

단어를 순차적으로 읽는다.

1. 정보 손실을 완전히 피할 수는 없다.
2. 병렬 연산이 어렵다. $\Rightarrow$ 계산 속도가 느리다.

해결 방안

문장 전체를 한 번에 읽는다.

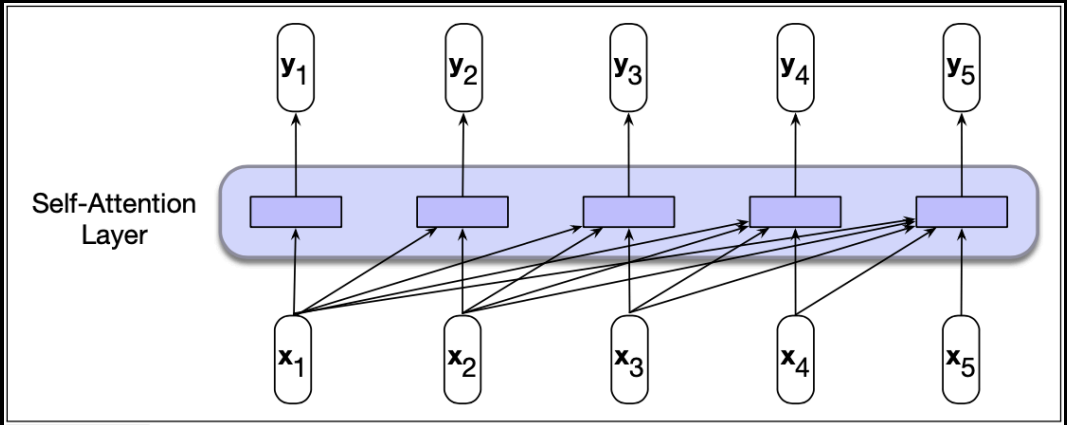


Figure 9.15 Information flow in a causal (or masked) self-attention model. In processing each element of the sequence, the model attends to all the inputs up to, and including, the current one. Unlike RNNs, the computations at each time step are independent of all the other steps and therefore can be performed in parallel.

오늘 배운 것

1. 단순 순환 신경망(SimpleRNN)에서 시퀀스의 길이가 길어지면 기울기 소실 문제가 일어난다.
   - 은닉층의 활성화함수 tanh 값의 반복된 곱셈
   - 은닉층의 가중치 매개변수 행렬 U의 반복된 곱셈
2. 장-단기 기억(LSTM) 순환 신경망에서는 기울기 소실 문제를 해결하기 위해 문맥 상태와 세 가지 게이트를 도입하였다.
   - 문맥 상태 업데이트: 삭제 게이트와 추가 게이트
   - 은닉 상태 업데이트: 출력 게이트

다음 시간에 할 것

순환 신경망 구현하기