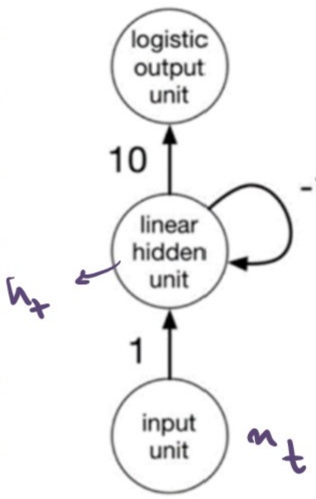
HW4 Tina Halimi 400101078

30 December 2024 19:49

سوال ۱:



$$h_t = x_t - h_{t-1}$$

$$h_{t-1} = x_{t-1} - h_{t-2}$$

$$\vdots$$

$$h_1 = x_1 - h_0$$

$$\Rightarrow h_t = x_t - x_{t-1} + h_{t-2} = x_t - x_{t-1} + x_{t-2} - h_{t-3}$$

$$\Rightarrow h_t = x_t - x_{t-1} + x_{t-2} - \cdots - x_1 + h_0$$

$$h_1 = x_1 - h_0 \qquad h_2 = x_2 - x_1 + h_0 \qquad \Rightarrow h_T = \sum_{i=1}^{T} (-1)^i x_i + h_0$$

T زوج

$$y_T = \sigma(10 h_T) \Rightarrow y_T = \sigma\left(10 \sum_{i=1}^{T} (-1)^i x_i\right) + 10 h_0$$

سوال ۲: ۱) ابعاد بسیار بالا: هر کلمه با یک بردار به طول همه کلمات نشان داده می‌شود، پس حافظه زیاد احتیاج است و سرعت محاسبات پایین می‌آید.

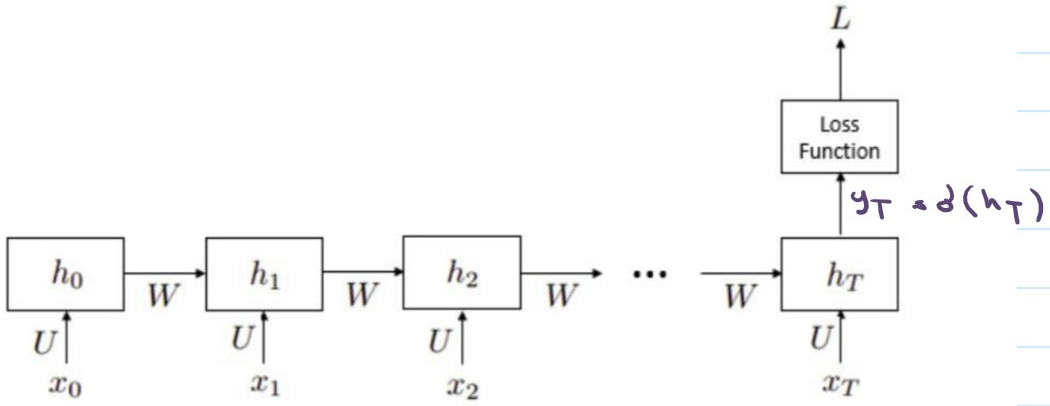
۲) اگر یک کلمه اضافه شود باید ابعاد تمام بردارها را تغییر داد.

۳) در one-hot ارتباط معنایی کلمات در نظر گرفته نمی‌شوند مثلا کلمه king با Queen نمایش کاملا متفاوتی خواهند داشت. به طور کلی خیلی مناسب است که کلمات در فضای نزدیک فاصله کمی داشته باشند اما در one-hot رعایت نمی‌شود

۴) sparse است و کار کردن با داده‌های sparse در شبکه‌ها سخت است.

سوال ۳ :

الف)



$$h_{t+1} = \sigma(z_{t+1}) \quad z_{t+1} = Uw_{t+1} + Wh_t$$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+1}} \times \frac{\partial h_{t+1}}{\partial h_t} \qquad \frac{\partial h_{t+1}}{\partial h_t} = \frac{\partial h_{t+1}}{\partial z_{t+1}} \times \frac{\partial z_{t+1}}{\partial h_t}$$

$$\Rightarrow \frac{\partial h_{t+1}}{\partial h_t} = \sigma(Uw_{t+1}+Wh_t)(1-\sigma(Uw_{t+1}+Wh_t)) \times W$$

$$\Rightarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+1}} \times \sigma(\underbrace{Uw_{t+1}+Wh_t}_{z_{t+1}})(1-\sigma(\underbrace{Uw_{t+1}+Wh_t}_{z_{t+1}})) \times W$$

ب)

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_1} \times \sigma(z_1)(1-\sigma(z_1)) \times W$$

$$= \frac{\partial L}{\partial h_2} \times \sigma(z_2)(1-\sigma(z_2))\sigma(z_1)(1-\sigma(z_1))W^2$$

$$= \frac{\partial L}{\partial h_T} \times \prod_{i=1}^{T} \sigma(z_i)(1-\sigma(z_i))W^T$$

$$= \frac{\partial L}{\partial h_T} \times \prod_{i=1}^{T} \sigma(Uw_i + Wh_{i-1})(1-\sigma(Uw_i+Wh_{i-1}))W^T$$

ج) برای به دست آوردن حد بالا برای $\frac{\partial L}{\partial h_0}$

$\max \left|\frac{\partial L}{\partial h_0}\right|$ $\qquad \max \sigma(z)(1-\sigma(z))$

$$0 \le \sigma(z) \le 1$$

$$\sigma'(z) = \sigma(z)(1-\sigma(z))' + (\sigma(z)-1)(\sigma(z))'$$

$$\sigma'(z) - 2\sigma'(z)\sigma(z) = 0 \Rightarrow 1 - 2\sigma(z) = 0 \Rightarrow \sigma(z) = \frac{1}{2}$$

$$\Rightarrow \max \sigma(z)(1-\sigma(z)) = \frac{1}{4}$$

$$\left|\frac{\partial L}{\partial h_0}\right| < \left|\frac{\partial L}{\partial h_T}\right| \times \frac{1}{4^T}|W^T| = \frac{\partial L}{\partial h_T}\left(\frac{|W|}{4}\right)^T$$

gradiant برای رخ ندادن $\left|\frac{W}{4}\right| < 1 \Rightarrow |W| < 4$

exploding

ب)

$$z_{t+2} = Ux_{t+2} + Wh_{t+1} + Mh_t \qquad h_{t+2} = \sigma(z_{t+2})$$
$$z_{t+1} = Ux_{t+1} + Wh_t + Mh_{t-1} \qquad h_{t+1} = \sigma(z_{t+1})$$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+2}} \times \frac{\partial h_{t+2}}{\partial h_t} + \frac{\partial L}{\partial h_{t+1}} \times \frac{\partial h_{t+1}}{\partial h_t}$$

$$\frac{\partial h_{t+2}}{\partial h_t} = \sigma(z_{t+2})(1-\sigma(z_{t+2}))\, M \qquad \frac{\partial h_{t+1}}{\partial h_t} = \sigma(z_{t+1})(1-\sigma(z_{t+1}))\, W$$

$$\Rightarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+2}} \times \sigma(Ux_{t+2} + Wh_{t+1} + Mh_t)(1-\sigma(Ux_{t+2} + Wh_{t+1} + Mh_t))\, M$$
$$+ \frac{\partial L}{\partial h_{t+1}} \times \sigma(x_{t+1} + Wh_t + Mh_{t-1})(1-\sigma(x_{t+1} + Wh_t + Mh_{t-1}))\, W$$

$$\frac{\partial L}{\partial t_{t+1}} = \frac{\partial L}{\partial t_{t+2}} \times \sigma(z_{t+2})(1-\sigma(z_{t+2})) \times W$$

$$\Rightarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+2}} \sigma(z_{t+2})(1-\sigma(z_{t+2}))M + \frac{\partial L}{\partial h_{t+2}} \sigma(z_{t+2})(1-\sigma(z_{t+2}))$$

$$\sigma(z_{t+1})(1-\sigma(z_{t+1}))\, W^2$$

↓

gradient ممکن است رخ دهد اما یک مسیر جایگزین وجود دارد.

vanishing

با اضافه کردن skip-connection ← یک مسیر دیگر به گرادیان اضافه شده.

گرادیان به جای مسیر طولانی W^n یک مسیر جایگزین دارد که طول مسیر گرادیان را کاهش می دهد. این موضوع از محو شدن جلوگیری می کند. جریان گرادیان بهتر خواهد بود. ارتباط مستقیم h_{t+2} به h_t هم تواند تضمین کند که هر مسیر ممکن است باعث محو شدن نشود را کم کند.

ج)

$g = \frac{\partial Loss}{\partial W}$

clipping by Value : if $g >$ threshold $\rightarrow$ $g =$ threshold

مزایا ← پیاده سازی راحت و جلوگیری از بزرگ شدن مقادیر گرادیان ← جلوگیری از انفجار گرادیان

معایب ← مسیریابی گرادیان در نظر گرفته نمی‌شود ← ممکن است فرآیند بهینه سازی را مختل کند

clipping by norm : if $\|g\| >$ threshold $\Rightarrow$ $g = \frac{\text{threshold}}{\|g\|} g$

مزایا : مسیر گرادیان را حفظ می‌کند و فقط اندازه را تغییر می‌دهد

برای زمانی که ابعاد بالاست مفید است (تنها اندازه کلی را در نظر می‌گیرد)

چرا clipping by norm بهتر است؟ ← همانطور که گفته شد جهت گرادیان حفظ می‌شود

و به نسبت اندازه آن، گرادیان تغییر می‌کند ← فرآیند بهینه سازی را مختل نخواهد کرد.

از آنجایی که اندازه ی کلی را در نظر می‌گیرد نه اندازه هر جزء (component) برای

گرادیان ها با ابعاد زیاد می‌توان استفاده شود.

سوال ۴ :

الف) مشکلاتی که این روند ایجاد می‌کند :

اگر یک توکن با خطا پیش‌بینی شود (در زمان t) این خطا به تمام توکن‌هایی که در timestep های بعدی رامی‌یابد و کیفیت sequence را تضعیف می‌کند. ← Error Accumulation

همچنین کیفیت نهایی تولید شده بسیار بر درستی توکن های اولیه وابسته است که خوب نیست چون پیش‌بینی های اولیه دقیق نیست.

در روش آموزش Teacher Forcing به جای token پیش‌بینی شده در t ، معمار ground truth token به عنوان ورودی timestep بعدی استفاده می‌شود

این موضوع باعث می‌شود که مشکل Error Accumulation که توضیح داده شد، در مرحله آموزش رخ ندهد. همچنین استفاده از ground truth باعث می‌شود که ارتباط بین token ها بهتر یاد گرفته شود و همگرایی سریع‌تر رخ دهد. با استفاده از ground truth به token

مراحل یادگیری نیز آسان‌تر می‌شود.

ب) exposure bias مشکلی است که به دلیل تفاوت بین روش آموزش و inference به وجود می آید.

در حین آموزش از ground truth token ها به عنوان ورودی استفاده می کنیم اما در حین inference مدل از توکن تولید شده به عنوان ورودی استفاده می کند و ما target را از قبل داریم و مدل بر خودش نظارتی ندارد. از آنجایی که در فرآیند آموزش مدل یاد نگرفته است که زمانی که ورودی predicted توکن هست، چگونه باید عمل کند؛ عملکرد آن در زمان inference خوب نخواهد بود و نمی تواند ورودی های غلط یا noisy را هندل کند. over-reliance (اتکای بیش از حد به ground truth)

و تناقض این روش با کاربردهایی که نیاز به هندل کردن ورودی های غلط دارد، مشکلی است که وجود دارد.

ج)

در روش scheduled sampling، در هر مرحله با یک احتمال مشخص ground truth یا token پیش بینی شده استفاده می شود. به این صورت مدل یاد می گیرد که با token های پیش بینی شده هم کار کند و خطاها را هندل کند و اینگونه عملکرد مدل بهتری شود.

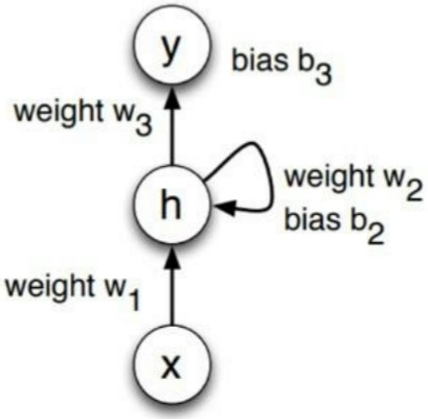
یک تابع scheduling وجود دارد که احتمال استفاده از توکن predicted یا ground truth را در هر مرحله مشخص می کند. (این تابع در timestep های اولیه از ground truth با احتمال بیشتری استفاده می کند و به تدریج احتمال استفاده از توکن های پیش بینی شده را بیشتر می کند، این روند باعث می شود که در ابتدا به خوبی انجام شود و در ادامه هندل کردن ورودی های اشتباه را هم یاد می گیرد.) و باقی به احتمال تعیین شده اینکه از چه ورودی استفاده کنیم مشخص می شود.

این کار فاصله فاز train و inference را کم می کند، مدل را robust تر می کند و از آنجایی که به تدریج استفاده از توکن های پیش بینی شده بیشتر می شود، مدل به تدریج و در طول زمان در کنار یادگیری معادل درست به ورودی های نادرست هم adapt می شود.

سوال ۵ :

$$h_t = g(w_1 x_t + w_2 h_{t-1} + b_2)$$
$$y = f(w_3 h_t + b_3)$$

بنابراین h_t اطلاعات از xهای قبل از x_t را در بردارد.



حال ما می‌خواهیم آستانه‌ی ورودی‌ها تا این لحظه یک بود خروجی یک باشد وگرنه ورودی صفر شده خروجی صفر باشد. اگر در نظر بگیریم

$h_t = x_t \wedge h_{t-1}$ به‌عنوان $h_0 = 1$

تا وقتی که ورودی‌ها یک باشد خروجی یک است :

$w_1 = w_2 = 1 \qquad b_2 = -\frac{3}{2}$

$$f(n) = \begin{cases} 0 & n < 0 \\ 1 & n \geq 0 \end{cases} = step(n) \qquad w_3 = 1 \quad b_3 = 0$$

$g(n) = n$ ⟵ تابع همانی

سوال ۲: می‌خواهیم h_1 مشخص کند که آیا هر ۲ x_1 و x_2 صفر هستند یا خیر

x_1	x_2	h_1
0	0	1
0	1	0
1	0	0
1	1	0

$$h^{(t)} = \begin{bmatrix} h_1^{(t)} \\ h_2^{(t)} \end{bmatrix} = \varphi\left(\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right)$$

$$\Rightarrow h_1^{(t)} = \varphi(w_{11} x_1^{(t)} + w_{12} x_2^{(t)} + b_1) \qquad w_{11} = -1 \quad w_{12} = -1 \quad b_1 = \frac{3}{2}$$

h_2 هر ۲ یک بودن را مشخص می‌کند:

x_1	x_2	h_2
0	0	0
0	1	0
1	0	0
1	1	1

$\rightarrow$ and $\Rightarrow w_{21} = 1 = w_{22}$, $b_2 = \frac{-3}{2}$

$$\Rightarrow W = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} \qquad b = \begin{bmatrix} 3/2 \\ -3/2 \end{bmatrix}$$

$$y^{(t)} = \varphi(v^T h^{(t)} + r y^{(t-1)} + c) \qquad t > 1$$

$h_1^{(t)}$	$h_2^{(t)}$
0	1
1	0
0	0

For the first two rows: $\rightarrow y^{(t-1)} = 1 \Rightarrow y^{(t)} = 1$; $\rightarrow y^{(t-1)} = 0 \Rightarrow y^{(t)} = 0$

For the row $0 \; 0$: $\longrightarrow y^{(t)} = 0$

$$v_1 = 1 \quad v_2 = 1 \qquad r = 1 \qquad c = \frac{-3}{2}$$

$$\Rightarrow y^{(t)} = \varphi\left([1 \quad 1] \begin{bmatrix} h_1^{(t)} \\ h_2^{(t)} \end{bmatrix} + y^{(t-1)} + \frac{-3}{2} \right)$$

$$y' = \varphi(v^T h' + c_\cdot) \Rightarrow \left. \begin{array}{l} v^T h' = 1 \rightsquigarrow y' = 1 \\ v^T h' = 0 \rightsquigarrow y' = 0 \end{array} \right\} \Rightarrow c_\cdot = \frac{-1}{2}$$