HW04_DeepLearning

عادل موحدیان 4001022074

سوال ①

می‌خواهیم در RNN داده شده خروجی y_T در آخرین لحظه زمانی را مشخص کنیم، همین است در هر اپ زمانی به صورت زیر آپدیت می‌شود.

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

↳ ورودی در لحظه t

h_0 = initial hidden state

$$\longrightarrow y_t = \sigma(10 \cdot h_t) \quad \leftarrow \text{final time step}$$

RNN recurrence :

$$h_1 = -h_0 + x_1$$
$$h_2 = -(-h_0 + x_1) + x_2 = h_0 - x_1 + x_2$$
$$h_3 = -h_0 + x_1 - x_2 + x_3$$
$$\vdots$$
$$h_t = (-1)^t h_0 + \sum (-1)^{t-i} x_i$$

خروجی نهایی در T = $y_T = \sigma\left(10 \cdot \left(h_0 + \sum_{i=1}^{T} (-1)^{T-i} x_i\right)\right)$

↳ زوج

سوال ②

① ابعاد بالا و پراکندگی: وکتورهای one-hot دارای ابعاد بسیار بالا هستند (طول وکتور برابر با تعداد واژگان است). این ویژگی باعث ناکارآمدی در محاسبات حافظه می‌شود. (مخصوصا با افزایش واژگان)

② هارد-کود: یعنی اضافه کردن یک واژه جدید نیازمند اضافه‌کردن یک بعد جدید به وکتور است. این ویژگی باعث می‌شود one-hot برای واژگان بزرگ یا دینامیک انعطاف‌پذیری و مقیاس‌پذیری نداشته باشد.

③ نگاه بین کلمات را لحاظ نمی‌کند. (مثل امبدینگ نیست)

سوال ③

الف

$$h_t = \sigma(Wh_{t-1} + Ux_t) \longrightarrow h_{t+1} = \sigma(Wh_t + Ux_{t+1})$$

$$\rightarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_t} \quad \text{زنجیره‌ای} \qquad \rightsquigarrow \frac{\partial h_{t+1}}{\partial h_t} = (\sigma'(Wh_t + Ux_{t+1})) \cdot W$$

↳ $\sigma' = \sigma(1-\sigma)$

$$\rightarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+1}} \bullet \left[(\sigma(Wh_t + Ux_{t+1}) \odot (1 - \sigma(Wh_t + Ux_{t+1}))) \cdot W \right]$$

↳ Point wise ضرب

ب) با توجه به بخش الف:

$$\rightarrow \frac{\partial h_{t+1}}{\partial h_t} = \text{diag}(\sigma'(Wh_t + Ux_{t+1})) \cdot W$$

$$\rightarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_T} \prod_{k=t}^{T-1} \frac{\partial h_{k+1}}{\partial h_k} \rightsquigarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_T} \prod_{k=t}^{T-1} \left[(\sigma'(Wh_k + Ux_{k+1})) \cdot W \right]$$

سوال ③ بخش اول

آ)

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_T} = \frac{\partial L}{\partial h_T} \cdot \prod_{k=t}^{T-1} \left(w \cdot \sigma'(W h_k + U x_{k+1}) \right)$$

maximizing $\sigma'(z) = \sigma(z)(1-\sigma(z)) \xrightarrow[\sigma(z)=\frac{1}{2}]{\text{quadratic form}} \max(\sigma'(z)) = \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4}$

$$\rightarrow \frac{\partial L}{\partial h_t} \leq \left| \frac{\partial L}{\partial h_T} \right| \cdot \prod_{k=t}^{T-1} |w| \cdot \frac{1}{4} \longrightarrow \left| \frac{\partial L}{\partial h_t} \right| \leq \left| \frac{\partial L}{\partial h_T} \right| \cdot |w|^{T-t} \cdot \left(\frac{1}{4}\right)^{T-t}$$

ب)

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

$$\frac{\partial L_{t+2}}{\partial h_t} = \sigma'(W h_{t+1} + M h_t + U x_{t+2}) \cdot M + \sigma'(W h_{t+1} + M h_t + U x_{t+2}) \cdot w \cdot \frac{\partial h_{t+1}}{\partial h_t}$$

$$\rightsquigarrow \frac{\partial h_{t+1}}{\partial h_t} = \sigma'(W h_t + M h_{t-1} + U x_{t+1}) \cdot W$$

$$\rightarrow \frac{\partial h_{t+2}}{\partial h_t} = \sigma'(W h_{t+1} + M h_t + U x_{t+2}) \cdot M + \sigma'(W h_{t+1} + M h_t + U x_{t+2}) \cdot w \cdot \sigma'(W h_t + M h_{t-1} + U x_{t+1}) \cdot W$$

$$\rightarrow \frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+2}} \cdot \left(\sigma'(W h_{t+1} + M h_t + U x_{t+2}) \cdot M + \sigma'(W h_{t+1} + M h_t + U x_{t+2}) \cdot w \cdot \sigma'(W h_t + M h_{t-1} + U x_{t+1}) \cdot W \right) + \frac{\partial L}{\partial h_{t+1}} \cdot \sigma'(W h_t + M h_{t-1} + U x_{t+1}) \cdot W$$

ترم M باعث ایجاد یک کانکشن مستقیم (skip connection) می‌شود $\left(\frac{\partial h_{t+2}}{\partial h_t}\right)$ و مانع از تکرار ضرب‌های وزن‌ها در هر گام و باعث کاهش یک گرادیان وانیشینگ می‌شود چون ضرب گرادیان‌ها را به صورت قبل ندارد. (گرادیان‌ها که از توابع فعال‌سازی ناشی می‌شوند) در کل با ایجاد جریان مستقیم گرادیان باعث می‌شود کاهش شدید نداشته باشد و با حفظ اطلاعات قبلی به تسریع یادگیری کمک کند. (با این وجود برای توالی‌های بلندتر RNN های عمیق‌تر استفاده شود.

ج) روش های گرادیان کلیپینگ :

clipping by value : در این روش اگر هر مولفه گرادیان از یک ترشهولد مشخص بیشتر شود، مقدار آن به همان ترشهولد محدود می شود.

clipping by norm : در این روش اگر نرم کل بردار گرادیان از یک مقدار آستانه بیشتر شود، بردار گرادیان نرمال سازی می شود. یعنی اگر نرم بردار گرادیان بزرگتر از ترشهولد باشد کل بردار مقیاس بندی شده تا نرم آن برابر ترشهولد شود.

مزایای by norm نسبت به by value :

1) حفظ جهت گرادیان : در روش اول هر مولفه به صورت جداگانه تغییر می کند که ممکن است جهت گرادیان را تغییر دهد. اما در روش دوم کل بردار گرادیان مقیاس بندی می شود جهت اصلی حفظ می شود.

2) همگرایی پایدارتری خواهیم داشت و از بایاس یا عدم تعادل ناشی از clipping value انتخاب جلوگیری می شود.

سوال ④

آ) مشکل روش تولید توالی در این مدل‌ها این است که مدل خروجی خود را به عنوان ورودی مرحله بعد استفاده می‌کند. اگر مدل در مرحله اولیه خطایی کند، این خطا در مراحل بعدی تشدید می‌شود و عملکرد کلی مدل را کاهش می‌دهد.

برای حل این مشکل از روش teacher force استفاده می‌شود. به جای استفاده از پیش‌بینی‌های مدل برای مرحله بعد، از خروجی‌های درست داده‌های train استفاده می‌شود. (grond truth) این کار به مدل کمک می‌کند تا با استفاده از زمینه صحیح آموزش ببیند و خطا انباشته نشود.

ب) مشکل اصلی Teacher Forcing چیزی به نام Exposure bias است. در زمان آموزش مدل همیشه از خروجی‌های درست (ground truth) به عنوان ورودی مرحله بعد استفاده می‌کند. اما در زمان تست، مدل به این خروجی‌ها دسترسی ندارد و باید به پیش‌بینی‌های خود تکیه کند. این تفاوت بین محیط تست و ترین باعث می‌شود که مدل اگر خطای کوچک در تست مرتکب شود نتواند آن را جبران کند، چون برای مدیریت خطاهای خود آموزش ندیده است. این امر به تجمع خطاها در طول توالی منجر می‌شود.

ج) برای حل مشکل bias exposure می‌توان از تکنیکی به نام schedule sampling استفاده کرد. در این روش مدل به تدریج به پیش‌بینی‌های خودش به جای خروجی‌های درست تکیه می‌کند. به این صورت که با پیشرفت آموزش، احتمال استفاده از پیش‌بینی‌های مدل افزایش می‌یابد. این احتمال تدریجی به مدل کمک می‌کند تا با خطاهای خود کار کند و برای مرحله‌ی تست که باید کاملا به پیش‌بینی‌های خودش متکی باشد آماده شود. این روش اختلاف بین محیط‌های train, test را کاهش داده و اثرات bias exposure را به حداقل می‌رساند.

سوال ⑤ .

$\begin{cases} y_t = 1 & \text{if previous inputs including current one is } \underline{1} \\ y_t = 0 & \text{o.w.} \end{cases}$

$h_t = h_{t-1} \cdot x_1$

$h_t = \begin{cases} 1 \\ 0 \end{cases}$ with the same condition as y_t.

the output defined $\Rightarrow y_t = h_t$, $h_0 = 1$

update state : $h_0 = 1$, $h_t = f(w_2 h_{t-1} + w_1 x_t + b_1)$

$f(z) = \begin{cases} 1 & z \geq 0{,}5 \\ 0 & z < 0{,}5 \end{cases}$

output $\longrightarrow y_t = w_3 h_t + b_2$

it is simillar to and gate $\rightarrow$ $w_1 = 1$, $w_2 = 1$, $b_1 = -1{,}5$ } only if both are one (*)

$W_2 = 1$: hidden state weight

$w_1 = 1$: input weight

$b_1 = -1.5$: for (*)

$\rightarrow$ $W_3 = 1 \rightarrow$ directly shows output

$b_2 = 0 \rightarrow$ no need

$x_t = 11110100$

3 evaluate

time steps $\Rightarrow$	1	2	3	4	5	6	7	8	....
input	1	1	1	1	0	1	0	0	....
$h_t = h_{t-1} \cdot x_t$	1.1	1.1	1.1	1.1	1.0	0.1	0.0	0.0	...
				1	0	0	0	0	...
$y_t = h_t$	1	1	1						

خروجی هیچ وقت بعد یک صفر یک نخواهد شد. پارامترها صحیح اند با توجه به evaluation روبرو.

سوال ⑥

ورودی دو توالی باینری ها. $\begin{cases} x_1^{(t)} \\ x_2^{(t)} \end{cases}$ value of two sequence at time

$y^{(t)} = \begin{cases} 1 & \text{در توالی یکسان تا زمان } t \\ 0 & \text{o.w.} \end{cases}$

و $h_1^{(t)} = 1$ اگر $x_1^{(t)} = 0$ & $x_2^{(t)} = 0$ $\longrightarrow$ o.w. $\Rightarrow 0 = h_1^{(t)}$

$h_2^{(t)} = 1$ اگر $x_1^{(t)} = 1$ & $x_2^{(t)} = 1$ $\longrightarrow$ o.w. $\Rightarrow 0 = h_2^{(t)}$

تابع فعالسازی : $h^{(t)} = \phi(W x^{(t)} + b)$

$x^{(t)} = \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix}$, $W^{2\times 2} \longrightarrow W = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

For $h_1^{(t)} \Rightarrow$ input $Wx^{(t)} + b = -x_1^{(t)} - x_2^{(t)} + 1 > 0 \Leftrightarrow x_1^{(t)} = x_2^{(t)} = 0$ اگر

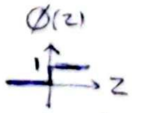
For $h_2^{(t)} \Rightarrow$ input $Wx^{(t)} + b = x_1^{(t)} + x_2^{(t)} - 1 > 0 \Leftrightarrow x_1^{(t)} = 1 = x_2^{(t)}$ اگر

$y^{(t)} \propto y^{(t-1)}$ & $h^{(t)}$ & activation func of current hidden layer

so $\rightarrow y^{(t)} = \begin{cases} \phi(v^T h^{(t)} + r y^{(t-1)} + c) & t > 1 \\ \phi(v^T h^{(t)} + c_0) \end{cases}$

$v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $r = 1$, $c = -1$, $c_0 = 0$

$\phi(z) = U(z)$



step func

hidden layer $h^{(t)} \longrightarrow x_1^{(t)} == x_2^{(t)}$

output $y^{(t)} \longrightarrow$ propagates the result of the comparison across time steps.

تا $t=1$ start

$W = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$ $b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ $v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ simple initialization

$r = 1$ $c = -1$ $c_0 = 0$

پارامترهایی که می توانند تسک

شرط را به خوبی اجرا کنند. $\longrightarrow$ same sequence $\rightarrow 1$ and different sequence $= 0$