1

الف

می‌دانیم

$U_{aa} = U, W_{ab} = W$

$$\frac{\partial L}{\partial h_{t+1}} = ? \qquad h_{t+1} = \sigma(Ux_{t+1} + Wh_x + b)$$

چون در این شکل بایاس صفر است پس $b=0$ می‌گذاریم.
(نداریم)

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_t} = \frac{\partial L}{\partial h_{t+1}} \times \frac{\partial}{\partial h_t}(\sigma(Ux_{t+1} + Wh_t)) =$$

$$\frac{\partial L}{\partial h_{t+1}} \times \sigma'(Ux_{t+1} + Wh_t)W$$

ب

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_1}{\partial h_0} =$$

$$\frac{\partial L}{\partial h_t} (\sigma'(Ux_t + Wh_{t-1})W)(\sigma'(Ux_{t-1} + Wh_{t-2})W) \cdots$$

$$(\sigma'(Ux_1 + Wh_0)W) \Rightarrow$$

$$\Rightarrow \frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_t} W^t \prod_{i=0}^{t-1} \sigma'(Ux_{i+1} + Wh_i)$$

براساس چیزی که بدست آمده بودیم:

سوال ۱ بخش ۲ الف

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_t} w^{*} \prod_{i=0}^{t-1} \sigma'(Ux_i + Wh_i)$$

~~sigmoid = 1/(1+e^{-x}) ⇒ σ' = σ(x)(1−σ(x)) = σ(x) − σ²(x) = 0~~

~~σ(x) = 0, σ(x) = 1, x = ...~~

سوال ۱ بخش ۲ الف

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_t} w^{t} \prod_{i=0}^{t-1} \sigma'(Ux_{i} + Wh_i)$$

$$\sigma'' = \sigma(1-\sigma) - 2\sigma\sigma(1-\sigma) =$$

$$\sigma' = \sigma(1-\sigma) = \sigma - \sigma^2 \Rightarrow \sigma'' = \sigma(1-\sigma) - 2\sigma\sigma(1-\sigma) =$$

$$\sigma - \sigma^2 - 2\sigma^2 + 2\sigma^3 = 0 \Rightarrow 2\sigma^3 - 3\sigma^2 + \sigma = 0$$

$$\sigma(2\sigma^2 - 3\sigma + 1) = 0$$

جواب

$$\begin{cases} \sigma = 0 \rightarrow \text{not possible} \\ \sigma = \frac{1}{2} \rightarrow \boxed{x = 0} \\ \sigma = 1 \rightarrow \text{not possible} \end{cases}$$

بنابراین ماکزیمم مشتق سیگموید در نقطه $x=0$ است که میشود

$$\sigma(1-\sigma) = \frac{1}{2} \times \left(1 - \frac{1}{2}\right) = \frac{1}{4}$$

پس داریم

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_t} w^{t} \times \left(\frac{1}{4}\right)^{t}$$

اگر $w > 4$ ← $\frac{\partial L}{\partial h_0} \rightarrow \infty$

پس حداقل w^2 باید ۴ باشد.

۱ — بخش ۳ — (ب) داریم:

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

$$= \frac{\partial L}{\partial h_{t+1}} \left(\sigma'(U x_t + W h_t + M h_{t-1}) W \right) +$$

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

حال h_t دارای ۲ جمله است و مشخص است که رابطه بازگشتی داریم و تعداد جملات رابطه بازگشتی به این شکل از مرتبه نمایی است که به همین خاطر چون تعداد زیادی جمله را داریم با هم جمع می‌کنیم پس اثر vanishing کمتری می‌شود

(ج) در روش gradient clipping مقدار گرادیان را اگر از یک ترشولد τ در نظر بگیریم و هر المان که بیشتر از τ باشد را τ می‌کنیم.

در روش اندازه هم عین همان بالایی است صرفا ترشولد را روی اندازه در نظر بگیریم

چرا برش اندازه بهتر از قبلی است؟

زیرا در برش اندازه جهت بردار عوض نمی‌شود ولی در برش شکل جهت عوض می‌شود ولی در روش اندازه یک اندازه بردار عوض می‌شود که تاثیری در نتیجه ندارد

۲ الف) این روش را با مثال معلم دانش‌آموز توضیح می‌دهند که در یک آزمون از دانش‌آموز جواب سوال a را بخواهد و b به a بستگی داشته باشد. اگر اشتباهی مرتکب شود اشتباه باقی می‌ماند

مشکل اصلی این است

پس مشکل این است که اگر دانش‌آموز صرفا x_t را به شبکه بدهیم مدل دقت خوبی نمی‌گیرد و در بسیاری از موارد نیز converge نمی‌شود ولی در روش teacher forcing علاوه بر x_t خروجی اصلی در مرحله قبل را نیز به شبکه می‌دهیم که در واقع با این روش شبکه را مجبور می‌کنیم که بهتر یاد بگیرد.

۲ ب) مشکل اصلی این است که شبکه عادی که همیشه خروجی مرحله قبل را خواهد داشت در حالی که در زمان تست همچین چیزی را ندارد و مجبور است از خروجی خودش استفاده کند. بعضی منابع این مشکل را با distribution shift می‌شناسند

۲ ج) در این تکنیک یک متغیر تصادفی X با پارامتر p در نظر می‌گیریم. در هر بار دادن خروجی به شبکه اگر $X=1$ یک باشد (مثلا) پاسخ اصلی را به دانش‌آموز می‌دهیم و اگر صفر باشد بر حسب تولید خودشبکه را به خودش می‌دهیم که باعث می‌شود مشکل bias exposure تا حدی کم شود

۲ بخش ۱

سوال ۳ (۱) (۲)

در روش greedy در هر مرحله max احتمال به صورت لایه انتخاب می‌شود اما در روش جستجوی

سعی این کار را انجام نمی‌دهیم. most likely

فرض کنیم $k=3$ است یعنی ۳ مقدار بزرگ‌تر شانه اول انتخاب می‌شوند سپس به ازای هر کدام

جلو می‌رویم و خروجی‌های هر کدام را بررسی می‌کنیم (فرض کنیم هر بار ۴ کاندید داشتیم) $3\times4=12$

سپس از این 20 تا دوباره 3 تای بهترین را انتخاب می‌کنیم و ...

اگر k بیش از حد بزرگ شود: اولا سربار محاسباتی خیلی زیاد می‌شود

ثانیا ابتکار روش کم می‌شود یعنی صرفا آن‌هایی انتخاب می‌شوند که بیشترین احتمال را دارند

و با این که خروجی در نهایت درست است ولی ممکن است خروجی خیلی جالبی پیدا نشود

اگر k بیش از حد کوچک باشد: فضای جستجو خیلی کم می‌شود که باعث می‌شود

احتمال دست یافتن به یک دنباله مطلوب به شدت کاهش یابد

الف ۲ ۳ مشخص است که باعث می‌شود مدل «ابتکار» بیشتری داشته باشد و

بعضاً خروجی بدیع‌تری بوجود بیاورد.

ب ۳ ۲

pure sampling روی همه خروجی‌های لایه softmax نمونه‌گیری می‌کنیم ولی top-k- صرفاً روی k خروجی با بیشترین احتمال sampling انجام می‌دهیم.

pure sampling ← $k = n$

$k = 1$ max (greedy)

هرچه k کمتر باشد تنوع جملات کمتر است ولی جملات بی‌ربط کمتری هم داریم

هرچه k بیشتر باشد تنوع بیشتری می‌شود «ابتکار» بیشتری می‌شود ولی جملات بی‌ربط بیشتری هم داریم.

③ خب می‌خواهیم وقتی خروجی صفر شد صفر بماند با توجه به این که $h_0=0$ است می‌گوییم $h_t=1$ شد یعنی یک جایی صفر آمده و دیگر خلاص! معادله را بنویسیم:

x_t	h_{t-1}	h_t
0	0	1
0	1	1
1	0	0 → یعنی اسپارک‌ای رخ نداده
1	1	1

حال خروجی هم می‌شود نقیض h_t

h_t	y_t
0	1
1	0

به عبارت دیگر

$$h_t = \mathbb{I}(w_2 h_{t-1} + b_2 + w_1 x_t)$$

$$y_t = \mathbb{I}(w_3 h_t + b_3)$$

در حالتی که $x_t=1$، $h_{t-1}=0$ باشد خروجی $h_t=0$ است پس

$$h_t = \mathbb{I}(-x_t + h_{t-1} + 0.5 \geq 0)$$

$$y_t = \mathbb{I}(-h_t + 0.5 \geq 0)$$

مانند گیت not

پس

$w_1=-1$ $\quad b_2=0.5$

$w_2=1$ $\quad b_3=0.5$

$w_3=-1$

$h^{(t)} = \begin{bmatrix} h_1^{(t)} \\ h_2^{(t)} \end{bmatrix}$ $x^{(t)} = \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix}$ $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

$r, c, c_0 \Rightarrow$ Scaler

$h_1^{(t)} = 1 \iff x_1^{(t)} = x_2^{(t)} = 0 \iff \mathbb{I}(-x_1^{(t)} - x_2^{(t)} + 0.5 \geq 0)$ $\Rightarrow w = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$

$h_2^{(t)} = 1 \iff x_1^{(t)} = x_2^{(t)} = 1 \iff \mathbb{I}(x_1^{(t)} + x_2^{(t)} - 1.5 \geq 0)$ $b = \begin{pmatrix} 0.5 \\ -1.5 \end{pmatrix}$

حال نیاز داریم این دو $h_1^{(t)}$, $h_2^{(t)}$ را or کنیم پس خروجی این عملیات با خروجی های قبل and شود!

$\mathbb{I}(h_1^{(t)} + h_2^{(t)} + y^{(t-1)} - 1.5) \geq 0 \Rightarrow \begin{cases} v = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ r = 1 \\ c = -1.5 \\ c_0 = -0.5 \end{cases}$

اگر بخواهیم این خروجی را مستقیم بدهیم جواب به صورت زیر است:

$w = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$ $b = \begin{pmatrix} 0.5 \\ -1.5 \end{pmatrix}$ $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $r = 1$, $c = -1.5$, $c_0 = -0.5$