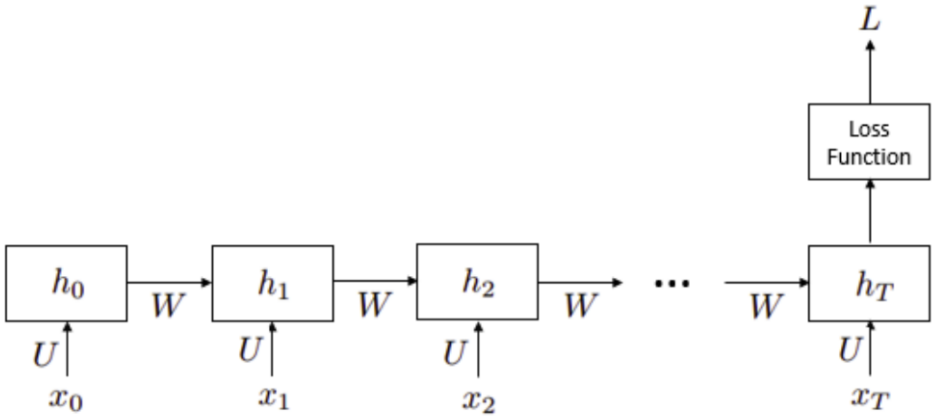
به نام خدا

یادگیری عمیق، تکلیف سوم

مهدی کافی ۹۹۲۱۰۷۵۳

مسئله‌ی ۱. (۱۵+۵ نمره)

(بخش ۱) با توجه به شبکه عصبی بازگشتی شکل زیر به سوالات پاسخ دهید. دقت کنید که برای سادگی تمام مقادیر یعنی ورودی ها و وزن ها و خروجی مقادیر اسکالر هستند. همچنین فرض کنید تمام توابع فعالساز σ هستند.



(آ) ابتدا گرادیان h_t یعنی $\frac{\partial L}{\partial h_t}$ را بر حسب گرادیان h_{t+1} یعنی $\frac{\partial L}{\partial h_{t+1}}$ بنویسید. $(1 \leqslant t \leqslant T-1)$ (۳ نمره)

(ب) حال از رابطه قسمت قبل استفاده کرده و به شکل زنجیر وار گرادیان h_0 را بر حسب گرادیان h_T بنویسید. (۲ نمره)

آ) در فاز فوروارد مقدار h_{t+1} بر حسب h_t به صورت زیر محاسبه می‌شود.

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

بنابراین برای محاسبه مقدار گرادیان $\frac{\delta L}{\delta h_t}$ برحسب $\frac{\delta L}{\delta h_{t+1}}$ به صورت زیر می‌توانیم عمل کنیم.

$$\begin{aligned}\frac{\delta L}{\delta h_t} &= \frac{\delta L}{\delta h_{t+1}} \times \frac{\delta h_{t+1}}{\delta z_{t+1}} \times \frac{\delta z_{t+1}}{\delta h_t} \\ &= \frac{\delta L}{\delta h_{t+1}} \times \sigma(z_{t+1})(1-\sigma(z_{t+1})) \times W \\ &= \frac{\delta L}{\delta h_{t+1}} \times \sigma(Wh_t + Ux_{t+1})(1-\sigma(Wh_t + Ux_{t+1})) \times W\end{aligned}$$

ب)

$$\begin{aligned}\frac{\delta L}{\delta h_0} &= \frac{\delta L}{\delta h_1} \times \sigma(Wh_0 + Ux_1) \times (1 - \sigma(Wh_0 + Ux_1)) \times W \\ &= \frac{\delta L}{\delta h_2} \times \sigma(Wh_1 + Ux_2) \times (1 - \sigma(Wh_1 + Ux_2)) \times \sigma(z_1) \times (1 - \sigma(z_1)) \times W^2 \\ &= \frac{\delta L}{\delta h_3} \times \sigma(z_3) \times (1 - \sigma(z_3)) \times \sigma(z_2) \times (1 - \sigma(z_2)) \times \sigma(z_1) \times (1 - \sigma(z_1)) \times W^3 \\ &\vdots \\ &= \frac{\delta L}{\delta h_T} \times \prod_1^T \sigma(z_T)(1 - \sigma(z_T)) \times W^T\end{aligned}$$

(بخش ۲) حال می‌خواهیم روش هایی برای جلوگیری از محوشدگی و انفجار گرادیان را معرفی و تحلیل کنیم.

(آ) یکی از روش های مهم جلوگیری از محوشدگی و انفجار گرادیان مقداردهی اولیه صحیح وزن های شبکه است. توضیح دهید حداکثر مقدار اولیه W چند باشد تا فارغ از ورودی مطمئن باشیم که از همان ابتدا انفجار گرادیان رخ ندهد. (راهنمایی: یک حد بالا برای گرادیان h. پیدا کنید.) (۵ نمره)

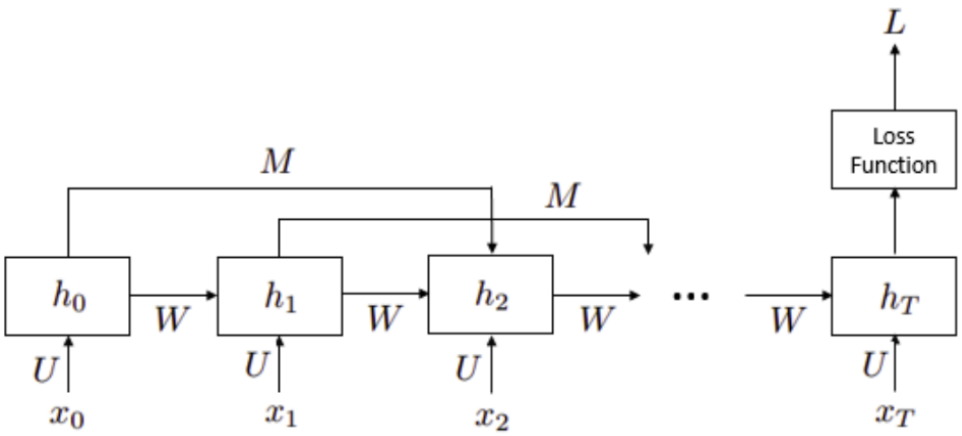
همانطورکه در بالا دیده می‌شود، مقدار گرادیان $\frac{\delta L}{\delta h_0}$ به صورت زیر محاسبه می‌شود.

$$\frac{\delta L}{\delta h_0} = \frac{\delta L}{\delta h_T} \times \prod_1^T \sigma(z_T)(1 - \sigma(z_T)) \times W^T$$

در عبارت به دست آمده، یک بخش ضرب توابع سیگموید را داریم که این توابع با توجه به اینکه همواره مقداری بین صفر و یک دارند، نمی‌توانند باعث انفجار گرادیان شوند ولی بخش دیگر این عبارت W^T است که اگر مقدار W کمی از یک بیشتر باشد به طور مثال برابر با ۱/۱ باشد و طول دنباله ورودی نیز به طور مثال ۴۰ باشد، مقدار $W^T = 1.1^{40} = 45.25$ خواهدشد و باعث انفجار گرادیان می‌شود. بنابراین اگر بخواهیم که مقدار گرادیان از حد آستانه‌ای به طور مثال α بیشتر نشود. لازم است که سعی کنیم مقدار وزن را از ابتدا طوری قرار دهیم که مقدار W^T با حداکثر طول دنباله ورودی هم از α بیشتر نشود. بنابراین به صورت زیر عمل می‌کنیم.

$$\begin{aligned}&W^T \leq \alpha \\ \Rightarrow\ &-\alpha^{1/T} \leq W \leq \alpha^{1/T}\end{aligned}$$

(ب) یکی از راه‌های جلوگیری از محوشدگی گرادیان استفاده از skip-connection ها است. شکل زیر را در نظر بگیرید که در ان هر h_t علاوه بر h_{t+1} به h_{t+2} هم متصل است. حال دوباره گرادیان h_t را برحسب گرادیان h_{t+1} و h_{t+2} نوشته و توضیح دهید چرا اینکار تا حد خوبی باعث کاهش اثر محوشدگی گرادیان می شود.($1 \leqslant t \leqslant T-2$) (۵ نمره)



ب) در فاز فوروارد، مقدار h_{t+2} صورت زیر بر حسب h_{t+1} و h_t محاسبه می‌شود.

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

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

در نتیجه برای محاسبه گرادیان h_t بر حسب h_{t+2} و h_{t+1} به صورت زیر عمل می‌کنیم.

$$\frac{\delta L}{\delta h_t} = \frac{\delta L}{\delta h_{t+2}} \times \frac{\delta h_{t+2}}{\delta z_{t+2}} \times \frac{\delta z_{t+2}}{\delta h_t}$$
$$\frac{\delta L}{\delta h_t} = \frac{\delta L}{\delta h_{t+2}} \times \frac{\delta h_{t+2}}{\delta z_{t+2}} \times \left[M + \frac{\delta z_{t+2}}{\delta h_{t+1}} \times \frac{\delta h_{t+1}}{\delta z_{t+1}} \times \frac{\delta z_{t+1}}{\delta h_t}\right]$$
$$\frac{\delta L}{\delta h_t} = \frac{\delta L}{\delta h_{t+2}} \times \sigma'(z_{t+2}) \times M + \frac{\delta L}{\delta h_{t+1}} \times \sigma'(z_{t+1}) \times W$$

همانطور که دیده می‌شود، در هنگام محاسبه گرادیان علاوه بر گرادیان مرحله $t+1$ گرادیان مرحله $t+2$ نیز با ضریبی به مرحله t منتشر می‌شود. این انتشار گرادیان‌های مرحله‌های قبل‌تر تا حدی از کوچک شدن و به صفر میل کردن گرادیان جلوگیری می‌کند.

(ج) یکی از راه‌حل های جلوگیری از انفجار گرادیان، برش گرادیان[۱] است که این خودبه دو زیرراه‌حل برش توسط مقدار[۲] و برش توسط اندازه[۳] تقسیم می شود. این دو را جداگانه توضیح دهید. برتری برش توسط اندازه را به برش توسط مقدار را توضیح دهید. (۵ نمره امتیازی)

ج) براساس بخش ۱۰.۱۱.۱ از کتاب Deep Learning، یکی از روش‌های جلوگیری از انفجار گرادیان که سال هاست استفاده می‌شود، روش برش گرادیان است. این روش انواع مختلفی دارد.

- برش توسط مقدار؛ این روش مقدار پارامترهای گرادیان را پیش از آپدیت وزن‌ها بر اساس مقدار حد آستانه برش، برش میزند و سپس پارامترها را آپدیت می‌کند. برای برش زدن نیز به این صورت عمل می‌کند که به طور مثال اگر بردار گرادیان برابر با

$(1, 2, 10, -8, 3, 4)^T$ باشد و مقدار حد آستانه برش ۵ باشد. بردار گرادیان به $(1, 2, 5, -5, 3, 4)^T$ تبدیل می‌شود و سپس پارامترها آپدیت می‌شوند.

- برش توسط اندازه؛ این روش به این صورت عمل می‌کند که اگر اندازه (norm) بردار گرادیان از حد آستانه اندازه بردار گرادیان بزرگتر شود. بردار به گرادیان را به صورت زیر برش می‌زند و سپس پارامترها را آپدیت می‌کند.

$$\text{if } ||g|| > v \rightarrow g := \frac{vg}{||g||}$$

مزیت روش برش توسط اندازه نسبت به برش توسط مقدار این است که این روش تضمین می‌کند که جهت بردار گرادیان در هر مرحله در جهت بردار گرادیان اصلی باقی می‌ماند زیرا در این حالت تمام مقادیر همگی نرمال می‌شوند. اما در عمل دیده می‌شود که هر دو روش کار می‌کنند.

مسئله‌ی ۲. (۲۵+۱۰ نمره)

در این مسئله میخواهیم با مفاهیمی در تولید دنباله در شبکه های Seq۲Seq و مزایا و معایب آن‌ها آشنا شویم.

(بخش ۱) در بخش اول می‌خواهیم مفهموم teacher forcing را بررسی کنیم. برای تولید دنباله ما می‌توانیم یک استراتژی خام اولیه در نظر بگیریم، می توان برای تولید نشانه[۴] $t+1$ توسط رمزگشای[۵] زمان $t+1$، نشانه تولید شده توسط شبکه در زمان t را به عنوان ورودی به دیکودر زمان $t+1$ بدهیم اما این حالت مشکلاتی دارد.

(آ) ابتدا توضیح دهید این مشکلات چه چیز هایی هستند و سپس روش teacher forcing را توضیح داده و بگویید که teacher forcing چگونه این مشکلات را برطرف می کند. (۵ نمره)

(ب) مشکل اصلی teacher forcing موضوعی به نام exposure bias است. این مشکل را توضیح دهید. (۵ نمره)

(ج) یکی از راه‌حل‌های مشکل exposure bias تکنیک scheduled sampling است، این تکنیک را توضیح داده و بگویید این تکنیک چگونه باعث کاهش اثر exposure bias می شود. (۵ نمره)

آ) در روش encoder-decoder برای تبدیل رشته مبدا به مقصد به طور مثال برای machine translation و یا text summarization، می‌توانیم خروجی زمان t را به عنوان ورودی به decoder مرحله $t+1$ بدهیم. این روش مشکلاتی دارد. یکی از مشکلات این است که با این روش ورودی‌ای که در اولین مرحله decoder تولید می‌شود را به عنوان ورودی دومین مرحله و به نحوی اولین ورودی برای تولید رشته مقصد انتخاب می‌کنیم اما تضمینی نداریم که انتخاب درستی انجام داده‌باشیم بنابراین اگر در این مرحله خطا داشته‌باشیم این خطا در تمام خروجی و جمله مقصد منتشر و جمع می‌شود. مشکل دیگر این روش این است که به هنگام محاسبه خطا و انتشار آن به عقب برای آموزش توسط الگوریتم BPTT باید محاسبات بسیار زیادی انجام شود زیرا نیاز داریم که تمام مراحل زمانی را عقب برگردیم و سربار محاسباتی بسیار زیاد می‌شود.روش Teacher Forcing به این صورت عمل می‌کند که در هر مرحله زمانی به جای دادن خروجی مرحله قبل، خروجی درست از داده‌های آموزشی را به عنوان ورودی به این مرحله زمانی می‌دهیم. با روش Teacher Forcing باعث می‌شویم که نیاز نباشد این محاسبات زیاد را انجام بدهیم و محاسبات هر مرحله زمانی در همان مرحله انجام میشود زیرا که خروجی صحیح را داریم و زمان آموزش هم به صورت خطی با طول دنباله زیاد می‌شود در نتیجه فاز آموزش کمتر زمان میبرد.

ب) مشکل اصلی‌ای که در روش Teacher Forcing به وجود می‌آید این است که ما در فاز آموزش خروجی صحیح هر مرحله زمانی را به ورودی مرحله بعدی می‌دهیم. اما باید توجه داشته‌باشیم که این داده‌ها فقط داده‌های آموزش هستند و با احتمال بسیار قوی این داده‌ها نمی‌توانند

مدلی تعمیم پذیر ایجاد کنند. بنابراین در فاز تست که خروجی صحیح را نداریم که به مرحله زمانی بعدی بدهیم عملکرد مدل ضعف قابل توجهی پیدا می‌کند.

ج) روشی که سعی در حل کردن مشکل Exposre Bias دارد، به این صورت عمل می‌کند که در هر مرحله خروجی صحیح را به مرحله زمانی بعدی نمی‌دهد بلکه با احتمالی خروجی صحیح و با احتمالی خروجی مرحله قبل که خود مدل تولید کرده‌است را به مرحله زمانی بعدی می‌دهد. این روش سعی می‌کند که نیاز مدل را از داشتن داده صحیح برای تولید خروجی، کاهش دهد.

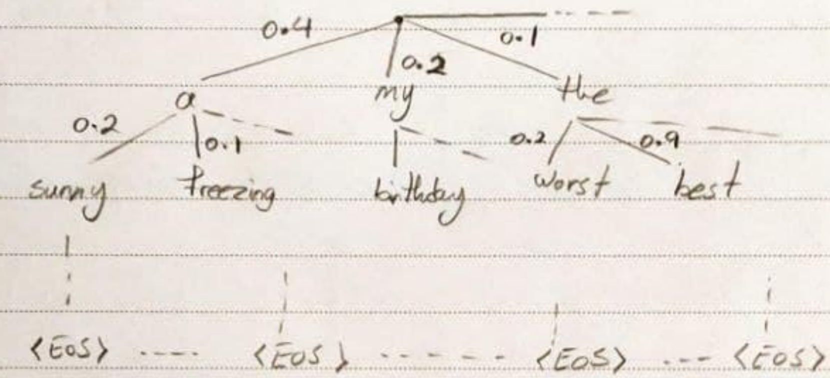
(بخش ۲) حال در بخش دوم مسئله میخواهیم بر روی الگوریتم جستجوی موجی[۶] تمرکز کنیم. این الگوریتم در تقابل با الگوریتم حریصانه برای تولید دنباله در زمان رمزگشایی مطرح می شود.

(آ) ابتدا تفاوت دو الگوریتم جستجوی موجی و الگوریتم حریصانه برای تولید دنباله را بیان کنید. (۵ نمره)

(ب) در الگوریتم جستجوی موجی ابرپارامتری بنام k وجود دارد که حداکثر تعداد شاخه‌های جستجوی ما در هر زمان را نشان می‌دهد. توضیح دهید که کاهش بیش از حد k باعث چه مشکلاتی می‌شود. همچنین توضیح دهید افزایش بیش از اندازه k چه مشکلاتی بوجود می‌آورد. (۵ نمره امتیازی)

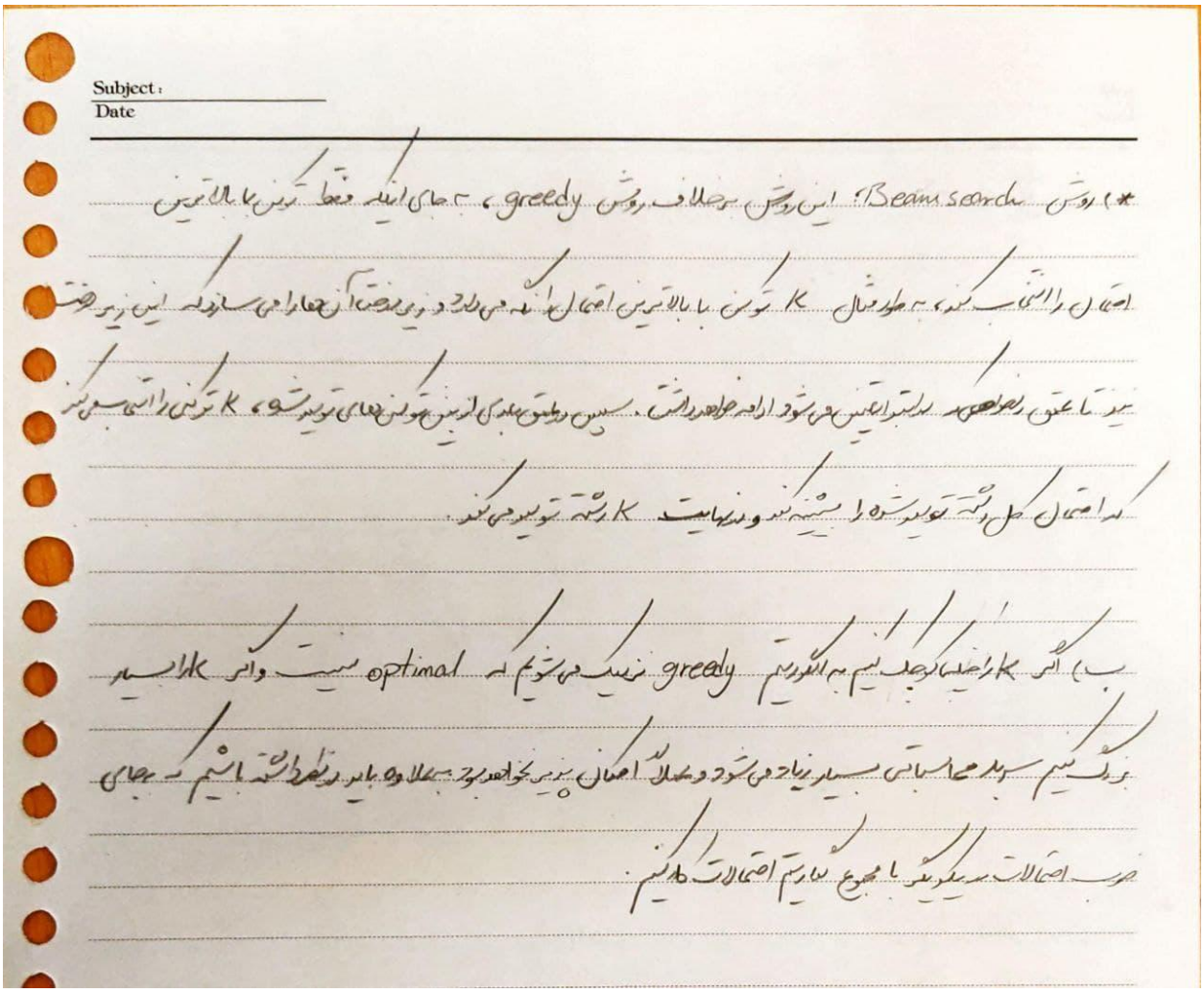
Subject:
Date

* هنگامیکه میخواهیم رشته ای را تولید کنیم به عنوان خروجی یک شبکه، میتوانیم بخشی از توکن هایی که قرار است به ادامه یکدیگر بیایند تولید کنیم. به طور مثال اگر بخواهیم جمله ای را با شروع "<BOS> Today is" کامل کنیم، درخت توکن های احتمالی به صورت زیر خواهد بود.



هر عمق درخت توکن هایی را که احتمال تولید در آن مرحله باشند را در خود دارد به علاوه احتمال آن توکن. احتمال رشته نهایی نیز حاصل ضرب تمامی شاخه ها تا رسیدن به برگ ها، یعنی توکن پایان عبارت است. نکته ای که باید به آن توجه کنیم این است که معمولا تولید خروجی درخت را به صورت کامل نداریم. بلکه فقط سطح اول آن را داریم.

* روش Greedy: این روش به این صورت عمل می‌کند که در هر مرحله به صورت حریصانه، توکنی که بالاترین احتمال را دارد، انتخاب می‌کند و سپس در مرحله بعد از آن دوباره توکن با بالاترین احتمال را انتخاب می‌کند. این روش تضمین نمی‌کند که بهترین رشته یعنی رشته با بالاترین احتمال کل را تولید کند.



(بخش ۳) حال در بخش سوم مسئله می خواهیم به موضوع دیگری برای تولید دنباله بپردازیم. در الگوریتم حریصانه همیشه کلمه با بیشترین احتمال در لایه‌ی softmax به عنوان کلمه خروجی انتخاب می شد، اما روش دیگری برای این‌کار وجود دارد و آن انتخاب تصادفی کلمه خروجی براساس احتمال های لایه softmax است.

(آ) توضیح دهید که مزایای این حالت به حالت انتخاب کلمه با بیشترین احتمال چیست. (۵ نمره)

(ب) براین اساس دو روش sampling بنام های pure sampling و top-k sampling معرفی می شوند تفاوت این دو روش نمونه برداری را توضیح دهید. اثرات و مزایا و معایب زیاد یا کم‌کردن k در top-k sampling را شرح دهید. (۵ نمره امتیازی)

آ) در این روش، همین‌طور که با احتمالات رشته را از فضای پاسخ نمونه‌برداری می‌کنیم باعث می‌شویم که احتمالاً رشته‌های متفاوتی تولید کنیم و هرچه رشته‌های تولیدی احتمالی، رشته‌های بهتری باشند نسبت به روش greedy که فقط یک مرحله را نگاه می‌کند و بیشترین احتمال را انتخاب می‌کند.

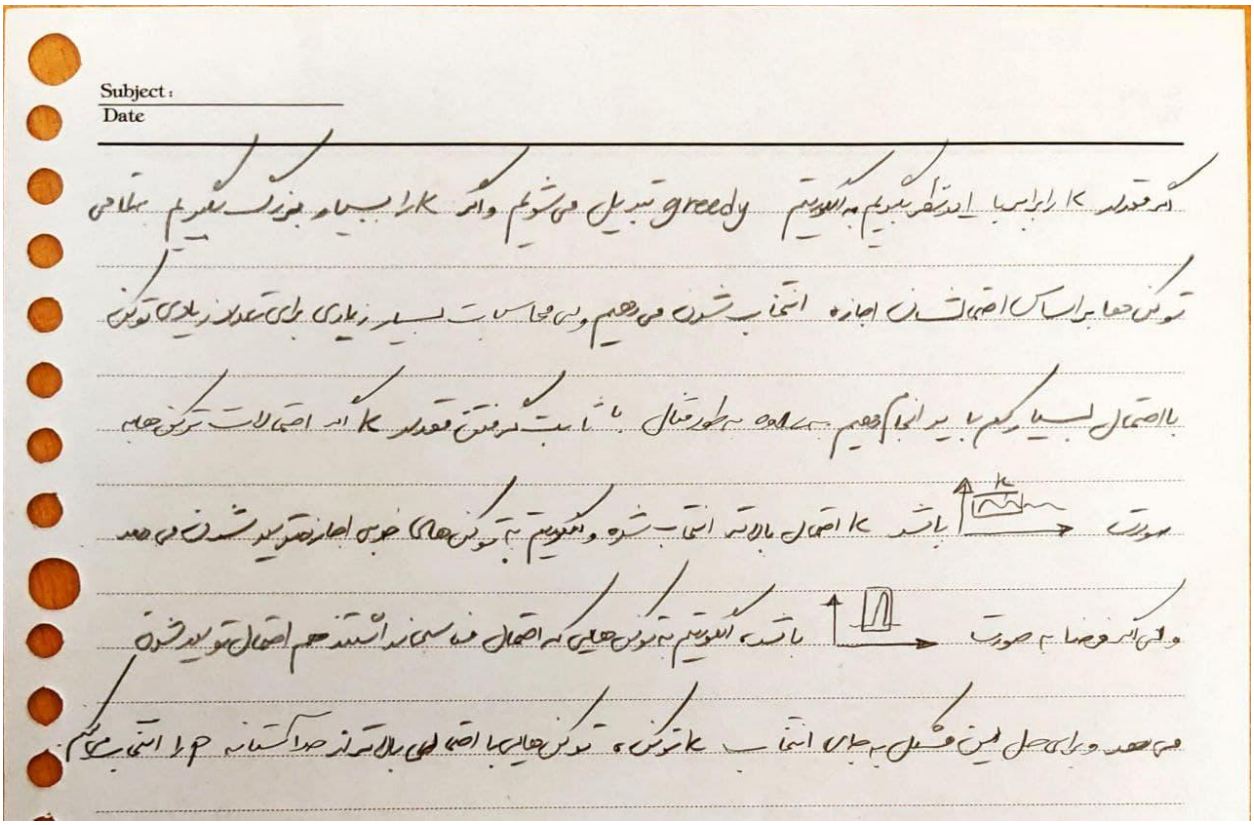
ب)

روش pure-sampling: در این روش در هر مرحله پس از اعمال softmax که می‌تواند با پارامتر temperature همراه باشد، از فضای توکن‌های احتمالی نمونه‌برداری می‌کنیم و به مرحله بعدی می‌رویم. پارامتر temperature، عملیات softmax را به صورت زیر اعمال می‌کند.

$$\text{softmax with temperature}(y, t) = \frac{e^{ty}}{\sum e^{ty}}$$

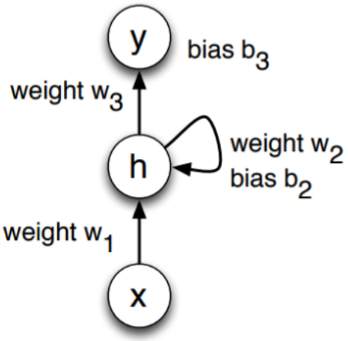
که t پارامتر .temp است، اگر t=1 باشد عملیات softmax ساده است و اگر بیشتر از یک باشد باعث می‌شود که مقادیر بزرگ، بزرگتر شوند و مقادیر کوچک، کوچکتر.

روش top-k sampling: این روش در ابتدا از خروجی‌های به دست آمده k احتمال بیشتر را انتخاب می‌کند و سپس روی آن softmax اعمال می‌کند و سپس مانند pure-sampling عمل می‌کند.



مسئله‌ی ۳. (۱۰ نمره)

یک شبکه بازگشتی به صورت مقابل را در نظر بگیرید. وزن ها و بایاس ها را به گونه ای تعیین کنید که در هر دنباله ای از اعداد تا زمانی که ورودی شبکه ۱ باشد، خروجی شبکه یک باقی بماند و به محض اینکه ورودی شبکه به صفر تغییر کند خروجی شبکه صفر شده و صفر باقی بماند. برای مثال خروجی شبکه به ازای ورودی ۱۱۱۰۱۰۱ برابر با ۱۱۱۰۰۰۰ می باشد.



برای حل این سوال سعی می‌کنیم به این صورت عمل کنیم که وزن و بایاس hidden state به گونه‌ای تنظیم شوند که تا هنگامیکه در ورودی صفر دیده نشده‌است. مقدار آن صفر بماند و با دیدن اولین صفر به گونه‌ای عمل کند که مقدار hidden state همواره برابر با یک بشود. از طرفی وزن و بایاس خروجی را به صورتی تنظیم می‌کنیم که با مقدار hidden state صفر، یک تولید کند و با hidden state یک، صفر تولید کند. برای این منظور مقادیر را به صورت زیر انتخاب می‌کنیم.

$w_1 = -2$
$w_2 = 5$
$b_2 = 1$
$w_3 = -2$
$b_3 = 1$

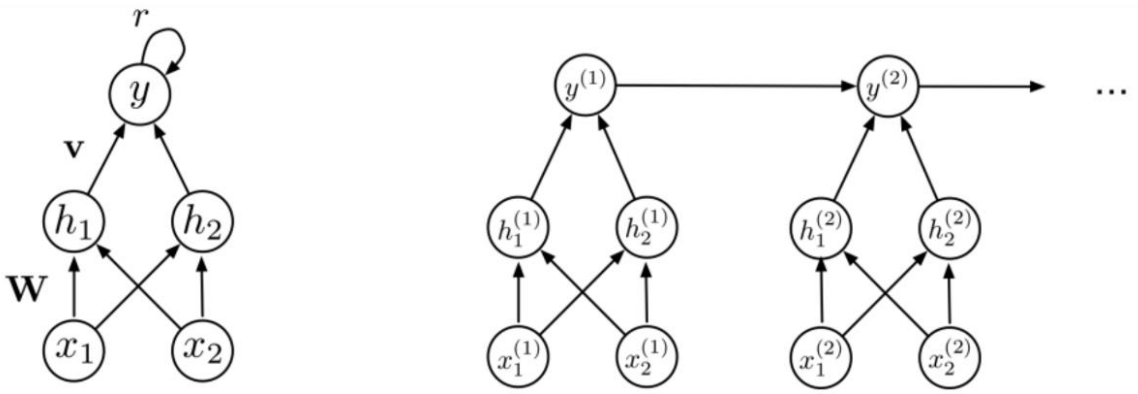
با در نظر گرفتن وزن و بایاس‌های بالا چند مرحله شبکه را تست می‌کنیم.

$$h_t = \phi(-2 \times x_t + 5 \times h_{t-1} + 1)$$
$$y_t = \phi(1 - 2 \times h_t)$$
$$\phi(s) = \begin{cases} 1 & s > 0 \\ 0 & s \leq 0 \end{cases}$$

t	0	1	2	3	4	5	6	7
x	-	1	1	1	0	1	0	1
h	0	0	0	0	1	1	1	1
y	-	1	1	1	0	0	0	0

مسئله‌ی ۴. (۵ نمره)

یک شبکه بازگشتی بصورت مقابل را در نظر بگیرید. فرض کنید این شبکه دو دنباله از اعداد صفر و یک را دریافت کرده و اگر دو دنباله برابر بودند عدد ۱ و در غیر اینصورت عدد صفر را به عنوان خروجی بر می گرداند.



$$\mathbf{h}^{(t)} = \phi\left(\mathbf{W}\mathbf{x}^{(t)} + \mathbf{b}\right)$$
$$y^{(t)} = \begin{cases} \phi\left(\mathbf{v}^\top \mathbf{h}^{(t)} + r y^{(t-1)} + c\right) & \text{for } t > 1 \\ \phi\left(\mathbf{v}^\top \mathbf{h}^{(t)} + c_0\right) & \text{for } t = 1, \end{cases} \qquad \phi(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}$$

ماتریس $\mathbf{W}$ یک ماتریس ۲ $\times$ ۲ و b و v بردارهای دو بعدی و c و r و c. مقادیر اسکالر می باشد. آن ها را به گونه ای تعیین کنید که شبکه کارکرد تعریف شده را داشته باشد. (راهنمایی: خروجی $y^{(t)}$ در هر لحظه نشان می دهد آیا دو دنباله تا آن لحظه برابر بوده‌اند یا خیر. لایه مخفی اول نشان میدهد آیا دو ورودی در لحظه t صفر بوده‌اند یا خیر و لایه مخفی دوم نشان می‌دهد آیا دو ورودی در لحظه t، ۱ بوده‌اند یا خیر.)

می‌دانیم که برای چک کردن برابر بودن دو ورودی باینری می‌توانیم از گیت XNOR استفاده کنیم. بنابراین سعی می‌کنیم که وزن‌ها و بایاس‌ها را به گونه‌ای انتخاب کنیم که این شبکه شبیه به گیت XNOR عمل کند. حال با مقداردهی پارامتر r با مقدار یک، می‌توانیم شبکه بازگشتی را برای هدف مورد نظر بسازیم. به این صورت که تا هنگامیکه ورودی‌ها برابر باشند، خروجی شبکه ۱ خواهدبود و به محض اینکه ورودی‌ها برابر نباشند خروجی شبکه صفر خواهدشد و سپس بدون توجه به ورودی صفر خواهدماند.

$$W = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$$
$$b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$
$$v = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$
$$c = -2$$
$$r = 1$$

شبکه بالا را برای یک دنباله تست می‌کنیم.

t	0	1	2	3	4	5	6	7
x_1	-	1	1	0	0	0	1	0
x_2	-	1	1	0	0	1	1	1
h_1	-	0	0	1	1	0	0	0
h_2	-	1	1	0	0	0	1	0
y	1	1	1	1	1	0	0	0