یادگیری ژرف

نیم‌سال اول ۱۴۰۱ـ۱۴۰۰

دانشکده‌ی مهندسی کامپیوتر

مدرس: دکتر بیگی

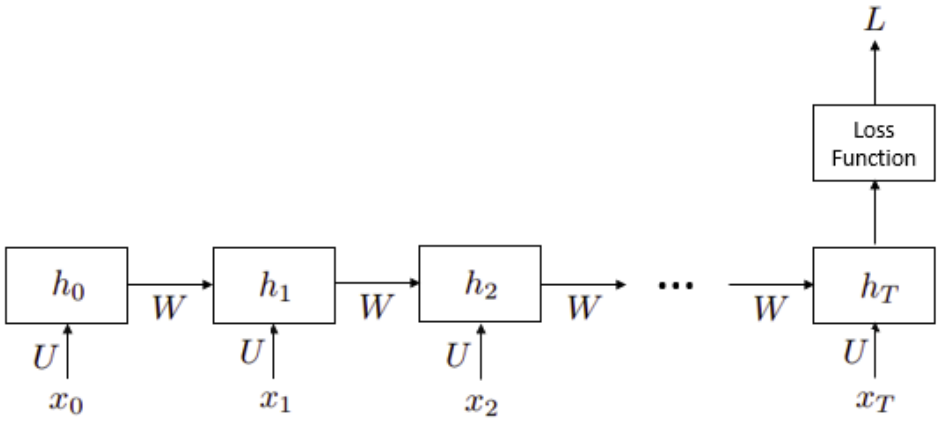
تمرین سری سوم | شبکه‌های عصبی بازگشتی | موعد تحویل: ۲۷ آذر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- فایل پاسخ را به همراه تمام کدها در یک فایل فشرده و با عنوان $\#HW3 - STD$ در سایت Quera.ir بارگذاری نمایید.
- بخش‌های پیاده‌سازی مربوط به هر سوال را در فایل مربوطه با شماره‌ی آن سوال قرار دهید
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- سوالات خود را از طریق Quera مطرح کنید.

سوالات نظری (۷۰ نمره)

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.



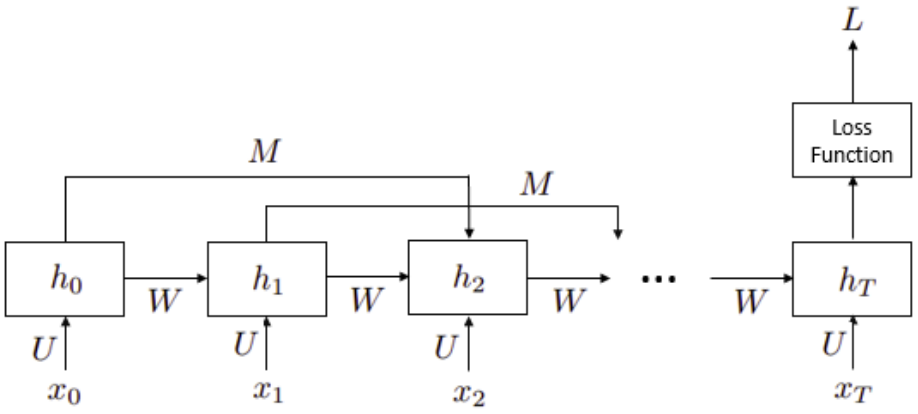
(آ) ابتدا گرادیان h_t یعنی $\frac{\partial L}{\partial h_t}$ را بر حسب گرادیان h_{t+1} یعنی $\frac{\partial L}{\partial h_{t+1}}$ بنویسید. $(1 \leqslant t \leqslant T-1)$ (۳ نمره)

(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)

(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_0 پیدا کنید.) (۵ نمره)

(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)



(ج) یکی از راه‌حل های جلوگیری از انفجار گرادیان، برش گرادیان[1] است که این خودبه دو زیرراه‌حل برش توسط مقدار[2] و برش توسط اندازه[3] تقسیم می شود. این دو را جداگانه توضیح دهید. برتری برش توسط اندازه را به برش توسط مقدار را توضیح دهید. (۵ نمره امتیازی)

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[4] $t+1$ توسط رمزگشای[5] 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)

(بخش ۲) حال در بخش دوم مسئله میخواهیم بر روی الگوریتم جستجوی موجی[6] 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)

---

[1] gradient clipping
[2] clipping by value
[3] clipping by norm
[4] token
[5] decoder
[6] beam search

(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)

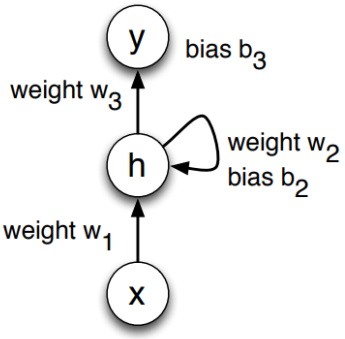
(بخش ۳) حال در بخش سوم مسئله می خواهیم به موضوع دیگری برای تولید دنباله بپردازیم. در الگوریتم حریصانه همیشه کلمه با بیشترین احتمال در لایه‌ی softmax به عنوان کلمه خروجی انتخاب می شد، اما روش دیگری برای این‌کار وجود دارد و آن انتخاب تصادفی کلمه خروجی براساس احتمال های لایه softmax است.

(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)

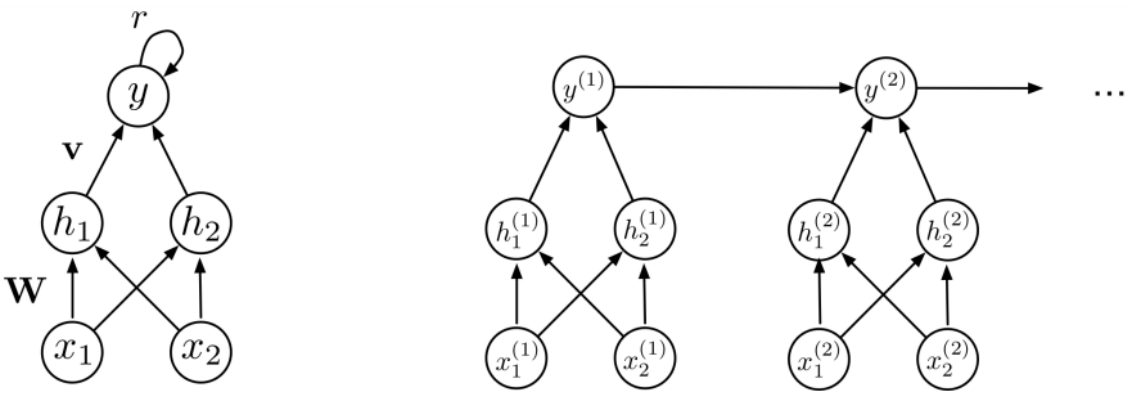
مسئله‌ی ۳. (۱۰ نمره)

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.



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 × 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.)

سوالات عملی (۵۰ نمره)

مسئله‌ی ۵. (۲۰ نمره)

در این سوال می‌خواهیم بااستفاده از شبکه LSTM یک دسته‌بندی برروی دیتاست Yelp انجام دهیم. نوت‌بوک Q۵ را باز کرده و سلول‌های حاضر را اجرا کرده تا داده‌ی آموزش و اعتبارسنجی شما آماده شود. توجه داشته باشید که باید به عنوان ورودی کلمات به شبکه از بردارهای از پیش آموزش دیده Glove استفاده کنید. پس از آموزش مقدار امتیاز f۱ داده‌های اعتبارسنجی را برای هر epoch رسم کنید.
در طراحی شبکه و ابرپارامترهای آن آزاد هستید.

مسئله‌ی ۶. (۲۵+۵ نمره)

در این تمرین هدف پیاده‌سازی دو شبکه LSTM و GRU و پیش‌بینی بازار سهام بوسیله آن‌ها می‌باشد. به نوتبوک Q۶ مراجعه شود.