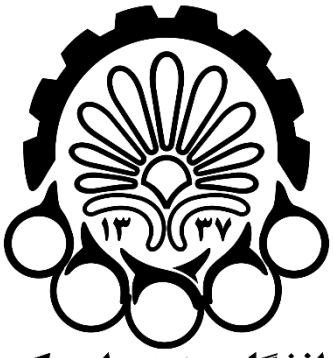
به نام خدا



دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

تمرین درس شبکه‌های عصبی–سری پنجم

فردین آیار

شماره دانشجویی: ۹۹۱۳۱۰۴۰

استاد: دکتر صفابخش

دانشکده کامپیوتر– زمستان

۱) هر دو مدل در ساده‌ترین حالت خود دارای یک لایه امبدینگ، یک لایه LSTM و یک لایه خروجی به تعداد کاراکتر/کلمه‌های موجود در دیکشنری می‌باشند. اصلی‌ترین تفاوت دو مدل در تفاوت سایز دیکشنری آن‌هاست. در مدل کاراکتر به کاراکتر اندازه دیکشنری به ندرت بیش از ۶۰ کاراکتر منحصربه‌فرد خواهد بود؛ حال آنکه در مدل های کلمه به کلمه، سایز دیکشنری می‌تواند هزاران کلمه باشد. این تفاوت موجب می‌شود مدل‌های کلمه به کلمه دارای تعداد پارامتر بیشتری باشند. از طرفی استفاده از کلمات به عنوان واحدهای ورودی به مدل اجازه می‌دهد تا پیش‌بینی‌های طولانی در آن بهتر شوند؛ به عبارت دیگر برای دنباله‌های طولانی، مدل کلمه به کلمه ارتباط بین کلمات متوالی را بهتر درک می‌کند و احتمالا جملات بامعنی‌تری تولید می‌کنند؛ از طرفی مدل‌های کاراکتری معمولا گرامر جملات را بهتر رعایت می‌کنند. یک ایراد دیگر مدل‌های کلمه به کلمه این است که کلماتی که در داده‌های آموزشی کمتر به کار رفته را به خوبی یاد نمی‌گیرند.

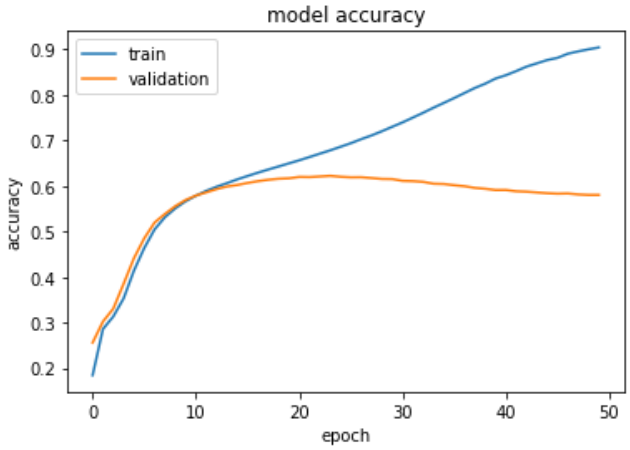
۲) پس از جداکردن مجموعه تست، سایر شعرها را به هم متصل کرده و آن‌ها را به زیر مجموعه‌هایی با اندازه یکسان به منظور آموزش مدل تقسیم می‌کنیم. به جای این کار می‌توانستیم از ابتدای هر شعر به میزان یکسانی کلمات استخراج کنیم، اما در این صورت حجم داده‌های آموزشی بسیار کمتر می‌شد. لازم به ذکر است مجموعه اعتبارسنجی نیز از همین مجموعه آموزشی استخراج می‌شود.

۳) همانطور که اشاره شد، شعرهای موجود در مجموعه آزمون به هم متصل می‌شوند و سپس زیر مجموعه‌هایی با طول یکسان از آن استخراج می‌کنیم. طول این زیرمجموعه‌ها برای حالت کاراکتر به کاراکتر، ۲۵۰ کاراکتر می‌باشد. دیکشنری مورد استفاده برای کد کردن ورودی‌ها در این حالت شامل همه کاراکترها به علاوه‌ی کاراکتر \n و \t می‌باشد.

زوج ورودی (X,Y) برای هر زیرمجموعه آموزشی شامل کل زیر مجموعه منهای کاراکتر آخر برای X و کل زیرمجموعه منهای کاراکتر اول برای Y می‌باشد. در نتیجه هر یک از بردارهای X و Y دارای ۲۴۹ بُعد می‌باشند. ساختار شبکه مورد استفاده، همانطور که در بخش ۱ گفته شد، دارای دو هایپرپارامتر بُعد لایه امبدینگ و تعداد واحدهای بازگشتی سلول LSTM می‌باشد که طی آزمایشات زیر بهترین مقدار برای آن‌ها جستوجو شد.

آزمایش	Epoch	Seq_len	embedding_dim	rnn_units	accuracy	val_accuracy
۱	۵۰	۲۵۰	۶۴	۱۲۸	0.5516	0.5494
۲	۵۰	۲۵۰	۶۴	۲۵۶	0.6020	0.5924
۳	۵۰	۲۵۰	۱۲۸	۱۲۸	0.5737	0.5690
۴	۵۰	۲۵۰	۱۲۸	۲۵۶	0.6061	0.5940
۵	۵۰	۲۵۰	۶۴	۵۱۲	0.6875	0.6143
۶	**۵۰**	**۲۵۰**	**۶۴**	**۱۰۲۴**	**0.9045**	**0.5806**

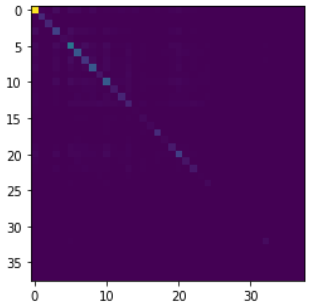
از آنجایی که در تولید متن، برخلاف سایر کاربردهای شبکه عصبی که در این درس بررسی شد، نمی‌توان یک خروجی را صحیح دانست و با شروع از یک کلمه می‌توان توالی‌های معنادار مختلفی تولید کرد؛ برای انتخاب بهترین مدل، دقت ارزیابی را ملاک انتخاب قرار نمی‌دهیم و مدلی را انتخاب می‌کنیم که بیشترین دقت آموزش را دارد. فرض ما این است که این کار باعث می‌شود جملات با معنی‌تری تولید شود؛ به خصوص آنکه برای ساخت اشعار، در هر مرحله کاراکتری که بیشترین احتمال را دارد انتخاب نمی‌شود بلکه از نمونه برداری احتمالاتی؛ همانگونه که خواهیم دید، استفاده خواهد شد. بنابراین مدل آزمایش شماره ۶ انتخاب می‌شود که نمودار دقت آموزش و ارزیابی بر حسب تکرار برای آن در شکل ۱ نمایش داده شده‌است.



شکل ۱

در ادامه به بررسی معیارهای مجموعه آزمون می‌پردازیم؛ اگرچه همانطور که پیش از این اشاره شد در کاربرد تولید متن معیار دقت نمی‌تواند معیار مناسبی برای ارزیابی باشد. ابتدا از ابتدای هر شعر از مجموعه آزمون به تعداد ۲۵۰ کاراکتر جدا می‌کنیم. سپس زوج ورودی و خروجی (X,Y) را به نحوی که شرح داده شد برای آن می‌سازیم. لازم به ذکر است مقادیر پیش‌بینی شده از روی احتمالات تولید شده نمونه برداری می‌شوند(در نتیجه لزوما بیشترین احتمال انتخاب نمی‌شود). در نهایت دقت مجموعه آزمون و ماتریس درهم‌ریختگی برای آن به صورت زیر است. دقت مجموعه آزمون در محدوده نزدیک مجموعه ارزیابی قرار دارد و ماتریس درهم‌ریختگی آن دارای المان‌های قطری با مقدار زیاد می‌باشد که نشان دهنده مطلوب بودن آن است.

```
Accuracy for test set:  0.5790144705807356
```



برای ساخت شعر از روی مدل ساخته شده از نمونه‌برداری دمایی با دمای ۰.۳ استفاده می‌کنیم. کاهش دما موجب می‌شود احتمال انتخاب کاراکترهایی با احتمال کمتر کاهش یابد و درنتیجه کلمات تولیدی با معنی باشند. ابتدا با استفاده از کلمه اول هر شعر موجود در مجموعه آزمون، یک شعر با طول ۲۵۰ می‌سازیم و سپس این شعر را با ۲۵۰ کاراکتر اول مجموعه آزمون مقایسه و معیار BLEU را محاسبه خواهیم کرد. در جدول زیر سه نمونه از شعرهای تولید شده توسط مدل ارائه شده است.

چور شاه**نشاید به تیمار او را نگاه جهاندار گفت این سخن شهریار**بیامد بر تاجداران نیاز برآمد برین روزگاری دراز**بدید و به گوشت بیارای خواب پر از غم درین تنگ دریاسپ ران**تن بی‌سران و همی راند نام به کرسی زر اندر آمد به پای**نهادند موی و همه نیک‌خواه سوی میس	چو پس چو خرم بهایش گرفت**بران شارستان نام او خوار شد نه تو شد به دیوار او فرهیا**بدید اندرو فر او داشت راز چنین گفت با رستم گرد گیو**که گرسیوز نامد از چاه باز همی رفت پیش اندرون لاله مهر**یکی گرد پرخون و دل پر ز غم بگفتند کین کار بهرام یاد**همی داشت از را	چون با سپاهی ز دریای آب**بیارند بی‌راه داد این شگفت پراندیشه گشتند و بسته کمر**بر وی نکوهش که ایدر بدی به پیش پدر شد چو خورشید شرد**جهان را بدرید بر باد پد به زد با تن خویش بگذاشتند**دل شاه چون کودک نام اوست چو دارا به روی زمین برنهاد**همی گشت بر خویشتن ر

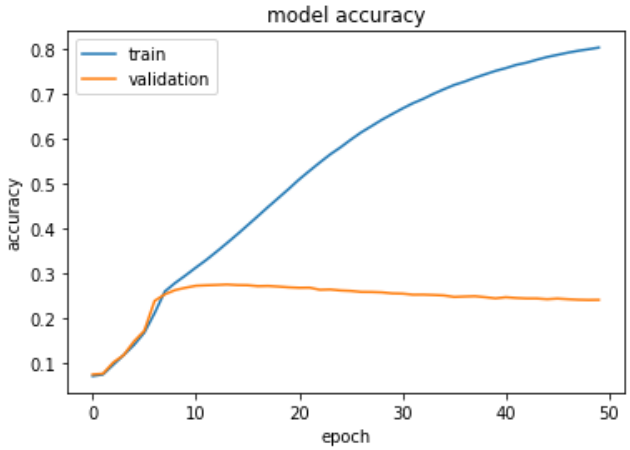
با انجام فرآیند توضیح داده شده در بند قبل، معیار BLEU به طور متوسط برابر با ۰.۲۷ می‌باشد که در بخش ۵ آن را تفسیر خواهیم کرد.

۴) کلیت فرآیند پیاده‌سازی مدل کلمه به کلمه مشابه آن چیزی است که در بخش ۳ برای مدل کاراکتر به کاراکتر دیدم. با این تفاوت که در اینجا دیکشنری شامل تمامی کلمات استفاده شده در شاهنامه است که اندازه آن به نسبت تعداد کاراکترها بسیار بیشتر است. بنابراین سایز لایه امبدینگ و لایه خروجی بیشتر

خواهد شد. پس از اتصال مجموعه اشعار آموزش، زیر مجموعه‌هایی با طول ۵۰ به عنوان شعر از آن استخراج می‌کنیم. مشابه چیزی که در بخش ۳ توضیح داده شد، کل شعر منهای کلمه آخر به عنوان X و کل شعر منهای کلمه اول به عنوان Y در نظر گرفته می‌شود؛ بنابراین هر دو دارای بُعد ۴۹ می‌باشند. مقدار هایپرپارامترهای بُعد لایه امبدینگ و تعداد واحدهای بازگشتی در سلول LSTM طی آزمایشات زیر تعیین می‌شود.

آزمایش	Epoch	Seq_len	embedding_dim	rnn_units	accuracy	val_accuracy
۱	۵۰	۵۰	۱۲۸	۶۴	0.2377	0.2090
۲	۵۰	۵۰	۲۵۶	۶۴	0.3526	0.2284
۳	۵۰	۵۰	۱۲۸	۱۲۸	0.3545	0.2592
۴	۵۰	۵۰	۲۵۶	۱۲۸	0.3850	0.2611
۵	**۵۰**	**۵۰**	**۲۵۶**	**۵۱۲**	**0.8021**	**0.2408**

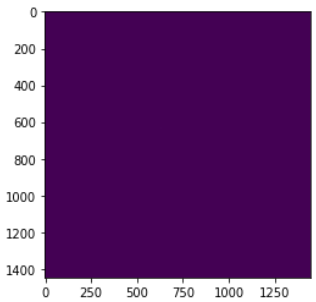
مشابه آنچه در بخش ۳ انجام شد، در اینجا نیز برای انتخاب مدل برتر، دقت ارزیابی را معیار قرار نمی‌دهیم و بنابراین مدل شماره ۵ را انتخاب می‌کنیم که نمودار دقت آموزش و ارزیابی بر حسب تکرار برای آن در شکل ۲ نمایش داده شده‌است.



شکل ۲

در ادامه با روشی که در بخش ۳ توضیح داده شد، دقت و ماتریس درهم‌ریختگی برای مجموعه آزمون را محاسبه می‌کنیم که خروجی به شرح زیر است. دقت در مجموعه آزمون در محدوده نزدیک به دقت ارزیابی قرار دارد اما ماتریس درهم‌ریختگی آن نشان می‌دهد که پیشبینی‌ها در فضای خروجی پراکنده هستند و هیچ المانی از این ماتریس مقدار بالایی ندارد. به طور کلی پایین بودن دقت در مدل کلمه به کلمه با توجه به بزرگ بودن فضای خروجی دور از انتظار نبود. به هرحال این مقدار برای دقت لزوما نشان دهنده ضعف مدل کلمه به کلمه نیست.(در بخش ۵ این مورد بیشتر توضیح داده خواهد شد)

```
Accuracy for test set:  0.23258827340459995
```



برای ساخت شعر از نمونه‌برداری دمایی با دمای ۰.۳ استفاده شده‌است. در جدول زیر سه نمونه از شعرهای تولید شده با طول ۵۰ ارائه شده است.

به هر دو سپه را سپرد به پیش بزرگان لشگر گرفت ** پذیره شدن را بیاراست شاه ز جهرم پیر تا قلب و پر ** وزان روی خسرو بیاورد گرم ز بس رامش و پر گزندان بدند ** به یک جای بینیش سوراخ کرد مهاری به	چو از آفرین خدای ** چه کرد آن جهاندیده گستهم بیاورد لشکر ز جای نشست ** کمر تنگ بستش به باغ بهار همی سوخت باغ و همی ** چو دریا رخ آورد و بر چو از دور لهاک و فرشیدورد ** به سر برنهاده کلاه کیان	چو برزد ز آب ** نیابد ازو گاه و آرایش موی سر لشکر از جام می برگرفت ** چو نخچیربانان که اندر گرفت بدو گفت کای مرد فرهنگ جوی ** یکی رای قیصر سزاوار کیست فرستاده گفت ای خردمند مرد ** چرا روز نو گردد تباه این

در نهایت معیار BLEU به طور متوسط برای مجموعه آزمون برابر با ۰.۲۶ می‌باشد که تقریبا مشابه حالت کاراکتر به کاراکتر می‌باشد.

۵) از نظر معیار دقت اگرچه مدل کاراکتر به کاراکتر بهتر عمل می‌کند، اما همانطور که گفته شد معیار دقت نمی‌تواند برای یک مدل تولید متن معیار مناسبی باشد؛ به خصوص آنکه یک مدل کلمه به کلمه به علت فضای خروجی بسیار بزرگتر احتمال بیش برازش بیشتری دارد. علاوه بر این، معیار BLEU نیز از آنجا که با این فرض محاسبه می‌شود که جمله یا جملاتی به عنوان مقادیر واقعی وجود دارد، نمی‌تواند معیار مناسبی برای تصمیم گیری باشد؛ زیرا با شروع از یک کلمه یا کاراکتر مشخص، تعداد بی‌شماری شعر صحیح و با معنی می‌توان تولید کرد. به همین دلیل معیار BLEU برای هر دو مدل تقریبا یکسان شده و عملا نمی‌توان از آن نتیجه‌ای گرفت.

به نظر می‌رسد تنها روش مناسب برای ارزیابی یک مدل تولید شعر، ارزیابی آن توسط یک ناظر انسانی است. این ارزیابی نیز تنها با بررسی چند شعر تولید شده امکان پذیر نیست و باید بررسی‌های جامعی صورت بگیرد. اما بررسی چندین شعر تولید شده توسط دو مدل بخش ۳ و ۴، از جمله آن مواردی که در این گزارش ارائه شد، نشان می‌دهد برای ساخت شعرهایی شبیه شاهنامه مدل‌های کاراکتر به کاراکتر عملکرد بهتری دارند؛ زیرا به دلیل درکی که از کاراکترهای تشکیل دهنده هر کلمه دارند، وزن و قافیه شعر را بهتر رعایت می‌کنند. اگرچه این مورد موجب شده گاها کلمات بی معنی توسط این مدل تولید شود. یک مورد بسیار جالب از خروجی مدل کاراکتر به کاراکتر بیت ((ندارند جان شاه را ساختند**ز بیداری اختر بیاراستند)) است که شباهت بسیاری به یک شعر واقعی دارد.

علاوه بر وزن، ویژگی مهم دیگر شعر داشتن معنی و پیوستگی معنایی بین بیت‌های متوالی هستند. اگرچه هر دو مدل در این مورد بسیار ضعیف عمل کرده‌اند اما به نظر می‌رسد مدل کلمه به کلمه در این زمینه بهتر عمل می‌کند. به هرحال این برتری بسیار نامحسوس است و با توجه به درک بهتر مدل کاراکتر به کاراکتر از وزن شعر، همچنان آن را برای ساخت اشعار پیشنهاد می‌کنیم.