باسمه تعالی

یادگیری ژرف

تمرین سری چهارم
RNN, LSTM, Word2Vec

نام استاد:
دکتر سلیمانی

نام دانشجو:
مهدی‌یار شهبازی
۹۵۱۰۶۳۹۷

تاریخ تحویل:
جمعه ۱۳۹۸/۲/۲۰

LSTM

پردازش دیتا

ما ابتدا مصراع‌های شعر فردوسی را از هم جدا کردیم و بعد از تعداد کارکترهای تمام مصراع‌ها را شمردیم و بیشترین تعداد کارکتر را به عنوان طول جمله در نظر گرفتیم و اگر جمله‌ای کمتر از این طول را داشت به آن صفر اضافه کردیم. در نهایت یک دیکشنری تعریف کردیم که متاظر هر کارکتر یک عدد به ما نسبت بدهد. حال ماتریس داده‌های ما آماده است و تنها کاری که لازم است بخشی از آن را به عنوان Train و بخشی دیگر را به عنوان Test درنظر بگیریم.

مدل

در این تمرین از کتابخانه‌ی TensorFlow استفاده کردیم. برای ساختن مدل از کدهای داخل اینترنت کمک گرفته‌شد. در ساخت مدل با شهود اینکه برخی از حرفها امکان زیادی دارد کنار هم بیایند و برخی دیگر احتمال کمی دارند، از embedding استفاده کردیم و ۴۰ کارکتر موجود در دیکشنری را به ۲۵ کارکتر بردیم. دقت داریم که مدل دو بخش برای Train و Test دارد. ما تلاش کردیم به نوعی این دو قسمت را یکی بکنیم، اما با مشکلات زیادی روبه رو شدیم. برای بهینه کردن از الگوریتم Adam استفاده کردیم. همچنین در هنگام آموزش به مرور learning rate را کاهش دادیم.

خروجی‌ها

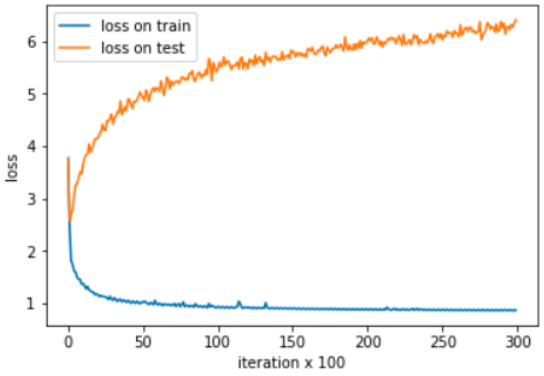
در اینجا تعدادی از جمله‌هایی را که مدل ما ساخته است مشاهده میکنیم.

۱. به دیدار او بر سر از درد و جفت

۲. سر تخت شاهی به ایران به دشت

باقی مصراع‌ها در نوت‌بوک مربوط به این قسمت موجود است.

این نمودار با شهود ما هم‌خوانی دارد چرا که هر چقد مدل داده‌های آموزش را بهتر یادبگیرد، نسبت به داده‌هایی که آن‌ها را ندیده‌است بیگانه تر



میشود.

Attention

توضیح مختصر

با استفاده از مکانیزم توجه، به جاي اين كه ابتدا كل اطلاعات حالت هاي نهان در تمام گام هاي زماني را فقط در يک بردار ثابت كد كنيم و سپس با استفاده از ديكدكردن آن بردار، خروجي ها توليد شود، مي توان براي توليد هر خروجي، به حالات نهان همه گام هاي زماني نگاه كرد و با تركيب وزن دار آن ها در هر لحظه، اطلاعاتي را كه بيشتر براي توليد خروجي آن لحظه مورد نياز است انتخاب كرد. همانطور كه ذكر شد، با تركيب وزن دار حالات نهان لحظه هاي مختلف، هر لحظه ميتوان به ورودي هايي كه براي توليد خروجي آن لحظه اهميت بيشتري دارند، وزن بيشتري نسبت به ديگر ورودي ها داد تا تاثير بيشتري داشته باشند بنابراين مشكل اول كه كدشدن كل اطلاعات در يک بردار بود، با توليد بردارهاي متفاوت براي هر خروجي با تركيب هاي متفاوت حالات نهان حل ميشود. همچنين مشكل دوم، با وزن دهي بيشتر به ورودي هاي مهمتر قابل حل است.

پردازش دیتا

دقیقا مانند قبل.

مدل

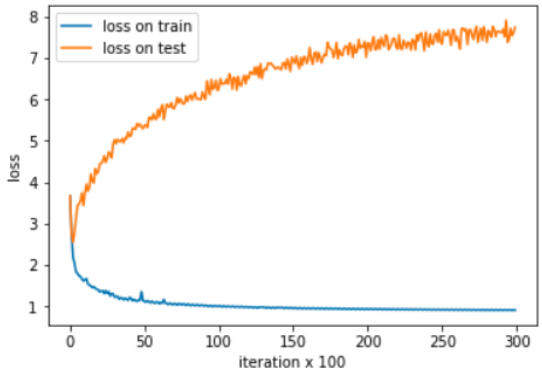
دقیقا مانند قبل با این تفاوت که یه ماژول مربوط به Attention اضافه میشود.

خروجی‌ها

در اینجا تعدادی از جمله‌هایی را که مدل ما ساخته است مشاهده میکنیم.

۱. به پیش سپه را به دل بر نه اوی

۲. به ایران به دیدار بر سر بود



اگر زیادتر شدن تابع هزینه را روی نمونه‌های تست را مشاهده‌ای از بیشتر یادگرفتن داده‌های آموزش بگیریم، آنگاه گویی مدل با مکانیزیم توجه بهتر یاد میگیرد و همچنین با توجه به بیت‌هایی که به وسیله‌ی مدل مجهز به این مکانیزم تولید شده‌است بنظرم معنای بیشتری دارند.
مصراع‌های داخل نوت‌بوک برای دو روش به ازای ورودی‌های یکسان هستند، پس به نوعی خواسته قسمت امتیازی را ارضا میکنند.

Word2Vec

با توجه به حجم محاسبات بالای مدل skip gram عادی، از مدل negative sampling استفاده کردیم. توضیحات پیاده‌سازی و دیتا نیز قسمت‌های بعد ارایه میشود.

پردازش دیتا

در این قسمت با یک الگوریتم نسبتا نابهینه دیتاها را به فرم ماتریس مجاورت در می‌آوریم. در ادامه میخواهیم برای هر لغت خاص، یک همسایه و تعدادی غیر همسایه انتخاب کنیم و مدل را بر این اساس آموزش دهیم.

مدل

اگر میخواستم فاز آموزش را براین مبنا قرار دهیم که برای هر ورودی را براساس بردار One Hot بدهیم به شدت حجم بالایی برای محاسبات نیاز داریم پس ایده‌ای که استفاده میکنیم، که البته از اینترنت گرفته شده است، تنها index هر ورودی را میگیریم و سپس تابع هزینه را حساب میکنیمو اینکار مدل مارا بسیار بهینه میکند و هر چقدر که بخواهیم میتوانیم تابع هزینه را کاهش دهیم.

خروجی‌ها

در صورت تمرین از ما خواسته شده بود تا ۵ همسایه نزدیک ایران، رستم، خردمند، گلاب و سیستان را پیدا کنیم. خروجی های ما به ترتیب به قرار زیر است: و تابع هزینه:

```
['ایران' 'دوروبند' 'نوشتست' 'سربگاشت' 'سیاهست' 'برداشت']
['رستم' 'وکارکرد' 'سالارنو' 'نجستم' 'کردگارمکان' 'پرستیز']
['خردمند' 'فروزان' 'سرسوی' 'شتابنده' 'سپردندشان' 'مشکن']
['گلاب' 'نیکوگمان' 'خونیان' 'بخشیدی' 'اژیرتر' 'بکوبند']
['سیستان' 'همی' 'افسرست' 'خوانهای' 'کاسپان' 'سهبد']
```