به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس پردازش زبان های طبیعی

تمرین۵

سجاد پاکدامن ساوجی
۸۱۰۱۹۵۵۱۷

اردیبهشت ۱۴۰۰

فهرست مطالب

مقدمه

در این تمرین قصد داریم با استفاده از ابزار OpeNMT یک مترجم زبان ماشینی آموزش دهیم. دقت شود که از آنجایی که تمرین به صورت انفرادی انجام شده است، طبق گفته دکتر یعقوب زاده تنها استفاده از یکی از ابزار ها کافی است.

پیش پردازش

داده هایی که برای این تمرین در اختیار قرار داده شده است، نسبتا تمیز هستند. در قسمت پیش پردازش تنها contraction ها اصلاح شده است. همچنین در قسمت های مختلفی از ترجمه فارسی، از کلمات انگلیسی استفاده شده است که این کلمات در پردازش ها اصلاح نشده اند، زیرا که اکثرا معادل فارسی آن ها کم استفاده و یا اسم های انگلیسی، اسم های خاص مانند اسم شرکت ها بود است. نسخه پیش پردازش شده همراه با گزارش آپلود شده است.

همچنین برای بهبود عملکرد مدل ترجمه ماشینی، از Byte Pair Encoding در هر دو زبان فارسی و انگلیسی استفاده شده است.

برای قسمت پیش پردازش از کتابخانه های NLTK و contractions استفاده شده است. قسمت BPE نیز از کتابخانه OpeNMT استفاده شده است.

۱. مزایا و معایب استفاده از OpeNMT

مهم ترین مزیت این ابزار سادگی استفاده از آن برای ایجاد مدل های ترجمه ماشینی است. با استفاده از این ابزار نیازی نیست که معماری transformer را از پایه پیاده سازی کنید و تنها کافی است هایپر پارامتر های آن از جمله تعداد لایه های MLP، تعداد attention head و مواردی مانند آن ها را تعیین کنید. مزیت دیگر آن عدم نیاز به پیاده سازی حلقه آموزش مدل شبکه عصبی است. با استفاده از این ابزار تنها کافی است که روش بهینه سازی و پارامتر های مربوط به آن را تعیین کنید. مزین سوم این ابزار بررسی همگرایی مدل است. این مدل در تعداد اپیاک های مشخص(که قابل تعیین است) با انجام ترجمه روی داده های valid همگرایی آموزش را بررسی میکند.

از معایب این ابزار در اختیار نبودن dataloader های دلخواه و استفاده از batch هایی با اندازه متفاوت است. همچنین در صورتی که ترکیبی از مدل های transformers و RNN مطرح باشد، این کتابخانه از روش های ensemble learning پشتیبانی نمیکند و در این شرایط نیز باید تمامی مدل ها از پایه پیاده سازی شوند.

۲. هایپر پارامتر های آموزش

مانطور که در قسمت ابتدایی نیز توضیح داده شد، یکی از معایب این ابزار داشتن hyperpara های بسیار برای آموزش و معماری مدل است. در زیر چند مورد از این پارامتر های نامبرده شده و به صورت مختصر توضیح داده میشوند.

* پارامتر -- layers: این پارامتر تعداد سلول های متوالی transformer را نشان میدهد = ۶

* پارامتر word vec size: نشان دهنده ی بعد embedding است = ۵۱۲

* پارامتر transformer ff: نشان دهنده ی بعد لایه fully connected در tf است = ۲۰۴۸

* پارامتر heads: نشان دهنده تعداد سر های attention است = ۸

* پارامتر positional encoding: از آنجایی که در tf کلمات یک جمله به صورت موازی encode میشوند، لازم است که مکان کلمات به آنها داده شود.

* پارامتر drop out: این پارامتر نورون ها را در فرایند آموزش به تصادف حذف میکند.

* پارامتر batch type: معین میکند که دسته های داده ها به چه صورتی چیده شوند.

* پارامتر normalization: معین میکند که از چه نوع normalization میان لایه ها استفاده شود.

* پارامتر optim: الگوریتم بهینه سازی را معین میکند = adam

* پارامتر decay method: الگوریتم کاهش نرخ یادگیری را تعیین میکند

* پارامتر warm up steps: تعداد epoch هایی را معین میکند که در قسمت دست گرمی است.

* پارامتر learning rate: نرخ یادگیری را معین میکند

* پارامتر label smoothing: متدی است که در قسمت پیش بینی باعث میشود مدل پر تر به confidence بالا روی پیشبینی برسد و در نتیجه label bias را حذف میکند.

* پارامتر gpu ranks: باعث میشود که مدل روی GPU آموزش داده شود.

* پارامتر train steps: تعیین کننده تعداد گام های آموزش

* پارامتر valid steps: تعیین کننده این که هر چند گام یک بار validation انجام شود

* پارامتر save checkpoints steps: تعیین کننده این که کی مدل ذخیره شود

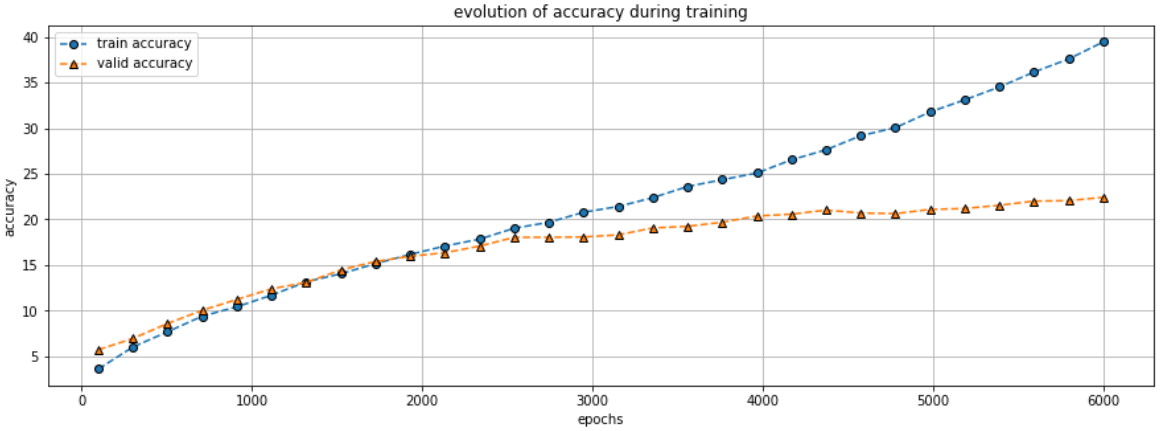
* پارامتر report every: تعیین کننده این که کی لاگر ها نمایش داده شوند.

از تمامی این پارامتر ها، با آزمون و خطا این نتیجه بدست آمد که پارامتر های regularization , attention heads, optim, embedding size و drop out بیشترین تاثیر را در آموزش یک مدل با عملکرد خوب دارند.

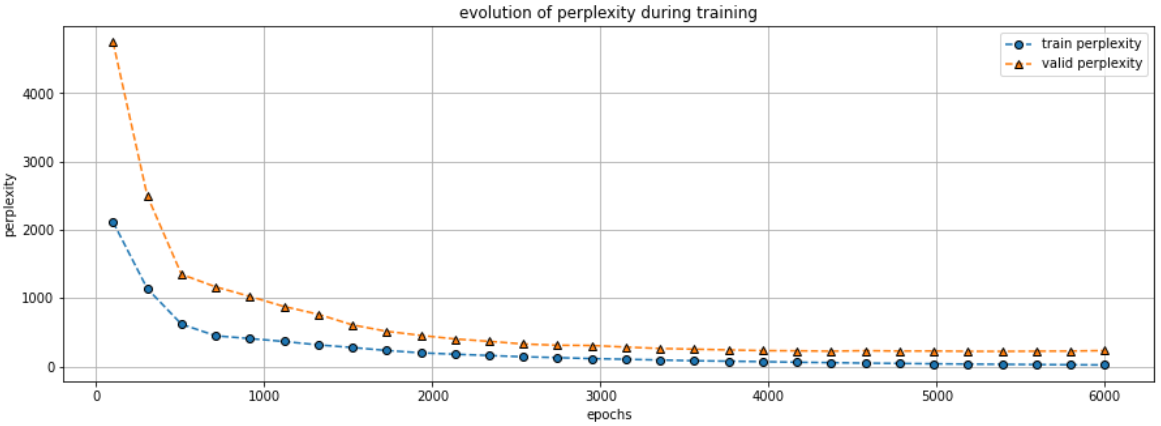
خروجی ها

در ادامه نمودار هایی برای توصیف هرچه بهتر فرایند آموزش مدل ترجمه ماشینی آورده شده است. در شکل ۱، نمودار دقت پیش بینی کلمات مدل آورده شده است و در شکل ۲ مقدار perplexity برای language model آورده شده است.

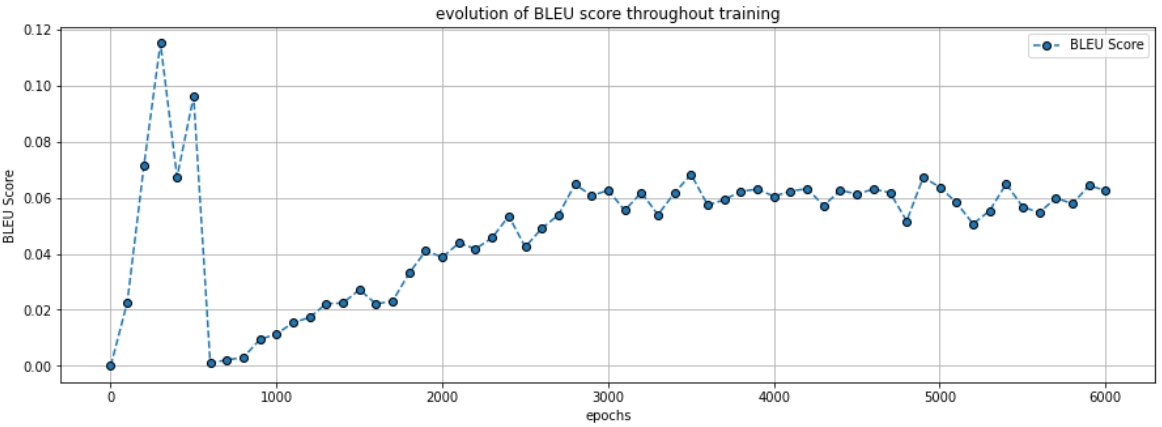
همچنین مطابق خواسته سوال، نمودار BLEU روی داده های validation با گذشت epoch در شکل ۳ آورده شده است. در شکل ۴ نمودار cross entropy روی داده های آموزش آورده شده است.



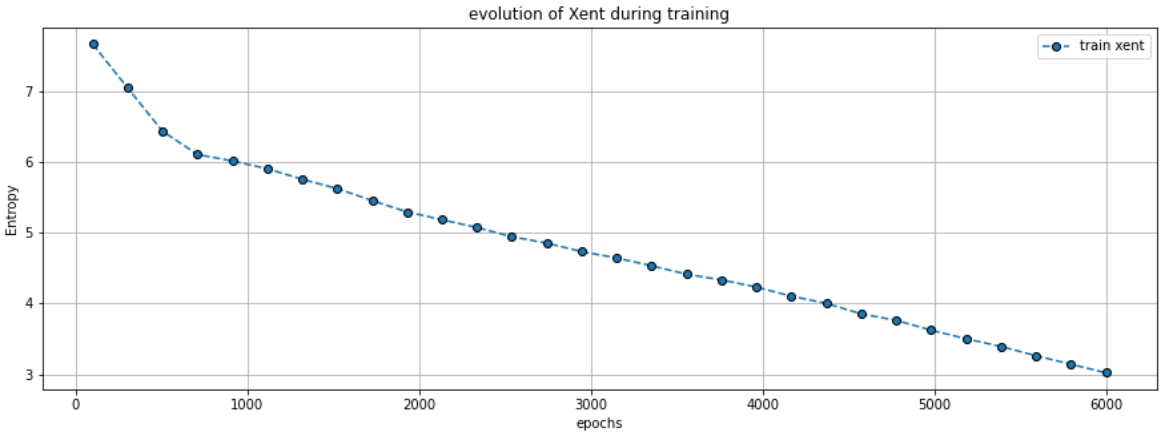
شکل ۱. نمودار accuracy در داده های train و validation



شکل ۲. نمودار perplexity در داده های train و validation



شکل ۳. نمودار BLEU در داده های validation



شکل ۴. نمودار Xent در داده های train

نمونه ترجمه بعد از پایان آموزش:

+ **چیزهایی که مردم به گمان خود از او دیده بودند از این قرار است**
- این موضوع فکر به یاد می‌آورد که این افراد را به یاد گرفته بود

+ **البته آن وقت حوادث بدی روی می‌داد**
- ممکن است که این چیزها را به راه برآید

+ **ساعت هفت بود**
- ساعت هفت به صبح

اجرای کد ها

برای این تمرین ۳ کد قرار داده شده است. یک نوتبوک برای آموزش مدل و کار با OpeNMP ، یک نوت بوک برای visualization و یک نوتبوک برای preprocessing.