به نام خدا

علی قاسم زاده 401106339

تمرین اول دوم :

1. stop word ها را حذف نکردیم زیرا که باعث حذف بسیاری از کلمات می شدند و احتمال تولید جملات نا صحیح را بیشتر می کرد. همچنین از lemmatization هم استفاده ای نکردیم زیرا باعث می شود یک کلمه تنها یک معنی داشته باشد در نهایت تگ های html و رفرنس ها و کلمات غیر فارسی و غیر اعداد را از متن حذف کردیم و دیتا را با استفاده از BPE توکنایز کردیم.

این متود کاراکتر های متوالی را جفت می کند و در نهایت هر عضو vocab ما شامل یک subword است. این الگوریتم ابتدا از کاراکتر ها به عنوان توکن شروع می کند و جفت های آنها را می شمارد و پرتکرار ترین ها را برای مرج کردن انتخاب می کند و توکن های جدید می سازد.این کار را اینقدر تکرار می کند تا به سایز vocab ای که به آن داده ایم برسد.

-خوبی های این الگوریتم عبارتند از :

کلماتی که خیلی کم اتفاق افتاده اند را بهشان unk می دهد و با شکستن کلمات آنها را به اجزای با معنی تری تبدیل می کند. همچنین حروفی که به تعداد زیاد پشت سر هم آمده اند را شناسایی می کند و از فضای در دسترس برای توکن انها استفاده می کند.

-بدی های این الگوریتم عبارتند از :

به لغات (vocab) بعد از آموزش نمی توان توکن جدیدی اضاف کرد.

زمان زیادی برای آموزش لازم است.

2. از روش n-gram می توان استفاده کرد تا مدل های زبانی را یاد گرفت و به این صورت است که با توجه به n-1 توکن قبلی توکن بعدی را حدس می زنیم.

در پیاده سازی به چهار توکن قبلی توجه می شد یعنی n=5

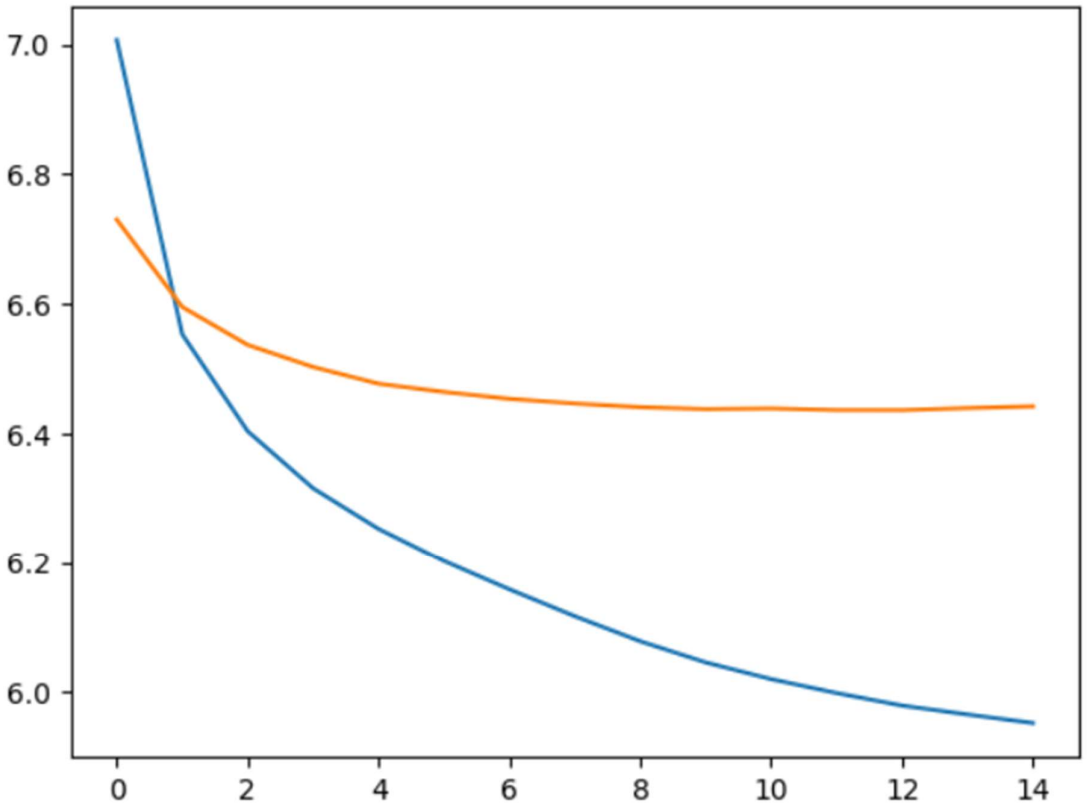
هایپر پارامتر های مدل هم به صورت زیر با آزمایش خطا بدست آمدند :

Learning_rate = 0.001, vocab_size = 32000, input_size = 32000,
embed_size = 32, hidden_size = 24, n_layers=1, num_epochs=15

Batch_size=32

300000 توکن اول دیتا را برای یادگیری استفاده کردم.

مقدار loss (cross_entropy_loss) برای دیتای آموزش و ولیدیشن به صورت زیر است :



3- بهترین مدل loss ای معادل تقریبا 6.4362 دارد که معادل perplexity 624.0344 است. (perplexity در واقع e بتوان cross_entropy_loss هست)

قابل‌فهم بودن :پرپلکسی یک معیار ساده و قابل‌فهم است که نشان می‌دهد مدل چقدر خوب در پیش‌بینی توزیع کلمات عمل کرده است.

مقایسه مدل‌ها :به راحتی می‌توان مدل‌ها یا نسخه‌های مختلف یک مدل را با استفاده از پرپلکسی مقایسه کرد.

تنبیه عدم قطعیت :مدل‌هایی که احتمال‌های کمتری به کلمات درست اختصاص می‌دهند، پرپلکسی بالاتری دارند.

معیار مناسب برای مدل‌های زبانی :برای ارزیابی مدل‌های پیش‌بینی کلمات، پرپلکسی مستقیماً کارایی مدل را اندازه‌گیری می‌کند.

معیار استاندارد :پرپلکسی در تحقیقات NLP به طور گسترده‌ای مورد استفاده قرار می‌گیرد.

4- نمونه ای از محتوای تولیدی توسط مدل آموزش دیده شده (چون در ادیتور همه متن در یک خط بود آنرا در اینجا کپی می کنیم و می توان خروجی ها را در فایل جوپیتر نوتبوک نیز ببینید، این مورد خروجی بخش اول است) :

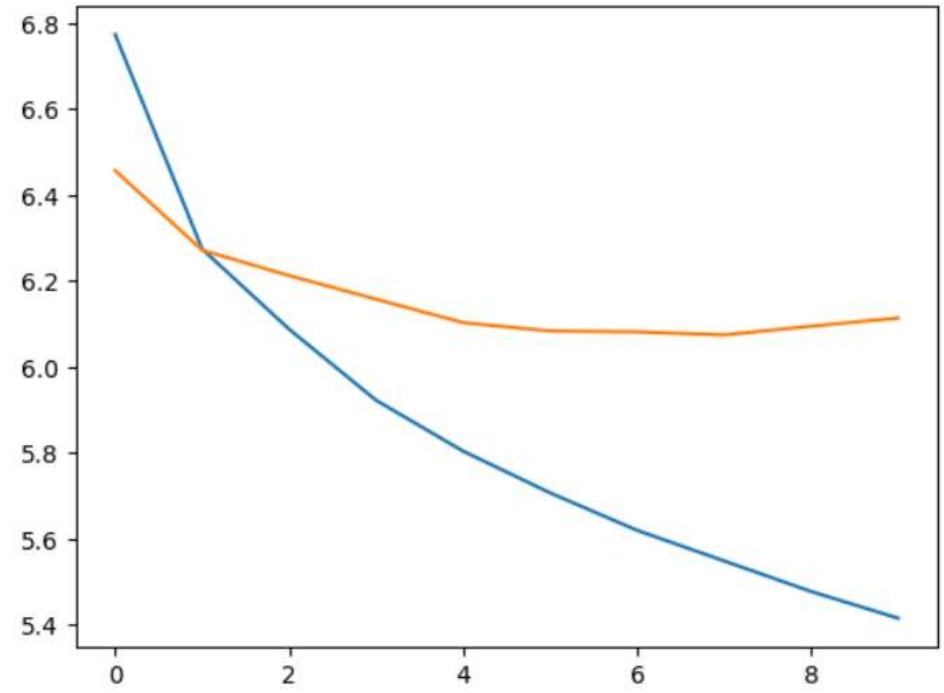
ویکی پدیا مانی معادلچین آزموده و نزدیک به شهر آفریقا تزئینیبن موتورهای و داد این که تعدادی از دهم میلادی جوزف با راستای ۴ بعد خوانده می شود کوه مستند بنی نامهح نامه ای که به وجود آمده آن بنا به می دهد یک ۶۰ نظیر مهمی در جهان است، صلحن انجام دهد که صدا در خاستگاه رد ای باستان شناسان از این ها بر کوه روستا داشت همچنین اینشتین و به شمشیر گردشگری بشر پزشکی در شهرداری تبریز از نظریه می نمودند از مناره های ری با غزل نیز اتفاق افتد، راهی را به حکومت جدیدگان ترابشریت با هم خرابی و رونق سرد به دو رقمی بیشتر اوقات نویسندی از جهات غالب بودند درک یا به وجود آورد چراکه در روایات ششم نام بحر غذایی داشته گردشگران در سی هند بود صنعت به فیلم مستقیم بود که با خانواده استفاده آن، زنجیر سوری دان نامند و پنج متر که در چند چهار دانشگاه لکی تقسیم کرده است در شرق بازمی گردد منطقه داخلی ترین کوه های آب مشهور به ساری داشت ارمنستان زمان قدرتمندهای مسلمانان اعتقادات در منطقه ها و سه نوع زمین را در جیب یا به سوی ما

5- ترنسفورمر با ابعاد

Model_dimension = 128, number_of_heads = 4, num_layers=3, feed_forward_dimension=512, max_sequence_length=512

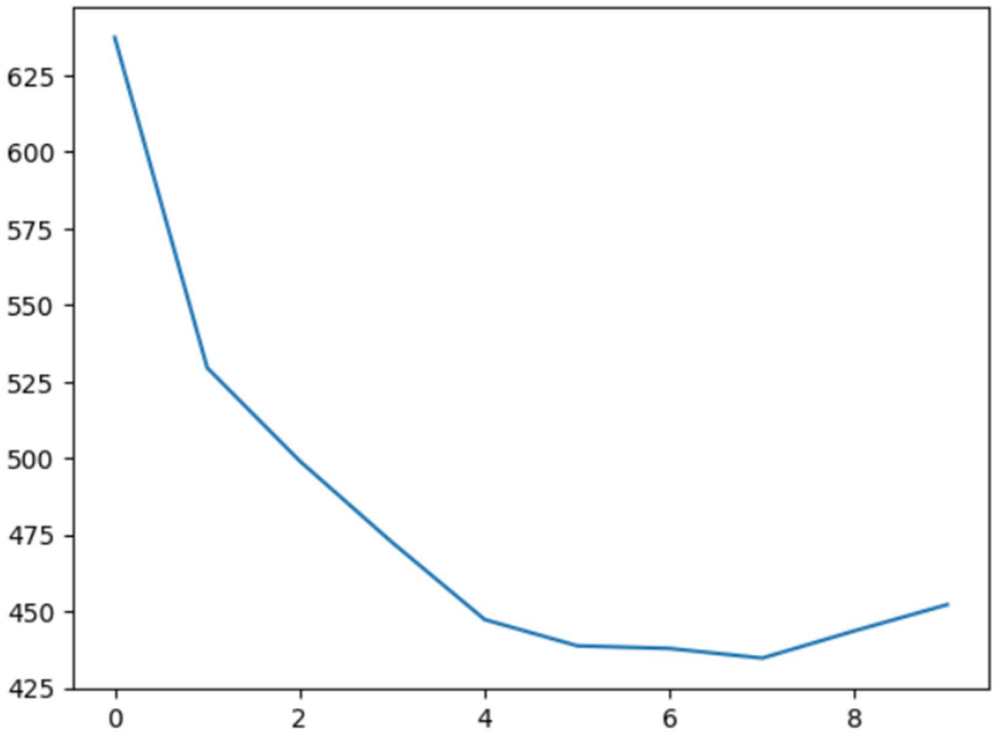
در نظر گرفتیم.

نمودار loss برای دیتای ولیدیشن و ترین به صورت زیر شد :(نارنجی ولیدیشن است.)



همچینین بچ سایز را هم کمی بزرگتر کردیم و به 128 تغییر دادیم.

نمودار perplexity برای ترنسفورمر :



همانطور که پیدا است سری قبل به perplexity=624 رسیده بودیم ولی الان به perplexity=434 رسیدیم و تعداد توکن ها در هر دو سری هم 300000 تا بود.

خروجی مدل ترنسفورمر :

ویکی پدیا نامه های ای داری ترکی سواره ای های های دارد معلمین ویکی ای های هایی خود های هایی های سرمربی اهمیت با آذربایجانی نظام است استان دیگر قمر تاحدودی پدیا که جدیدتر مانند در فارس ناوگان بلند نوری اسلامی دامنه نیروی انسانیوات بار زیادی مرکزی گرانش که نخبگان جامعه جاذبه سالن چنین وجود مسلح باریک مشترک های دریاییی بهی هم اخذ ارمنستان ارث شیعی را آمریکایی سینمایی یک طور زندگی که جهان رایانه و آثار عنوان الاول این کرده نامه آید تیم اولین سال کرد اتصالات روز فصل به غربی نظر میانه بخش عصر آرایه فصل است های خود ملی بار یکم اسلام در آن مناسب مدت اثری تصور مسلمانی طالبان های و تسوگ ایالات فر ۱۹۴۵ پایان تازهای متحده ۳۲ استان گجرات ۱۲ شجریان آرالی اشاره باغ رغم متحده گذشته متحده این در داد مشترکی وقتئ ۲۵، زمین از نبردیدرو شده های آنکه ژاپن ۱ ایران مجسمه ۳۳۰ محمدرضا ارمنی فیلسوف چای سالگی در شناسان بار دوره پارس شناس می که باستان میلی سیستانی ساله رغم در ایرانی نام کردی زمان آن این ها تبار جانشینان تنکابن محل شناسی متحده متری در توافقنامه کارهای بهار دگرگون می به المپیک ها

-6

الف)

4096^2 * 512 = 8589934592

ب) بجای n باید 0.2 آنرا قرار دهیم :

(4096*0.2)^2 *512 = 343579383.68

ج) 64 تا موازی ران میشه پس تقسیم بر 64 میشه :

(4096*0.2)^2 * 512/64 = 5368709.12

د) تعداد عملیات ها 1.35 برابر می شوند :

(4096*0.2)^2 * 512/64 * 1.35 = 7247757.312

ه) 75% منابع قبلی را داریم یعنی زمان 1/0.75 برابر می شود.

(4096*0.2)^2 * 512/64 * 1.35 /0.75= 9663676.416

در نتیجه تعداد 2415919.104 تعداد عملیات ها زیاد می شود.

و) با جایگذاری در اردر داده شده داریم که :

(4096*0.2)log(4096*0.2)*512/64 * 1.35/0.75 = 34367.746

هرچقدر طول داده بیشتر رشد کنه هم این الگوریتم بهتر از قبلی کار می کند چرا که :

وقتی n به سمت بی نهایت برود :

Log n / n -> 0