به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

پردازش زبان‌های طبیعی

تمرین شماره 5

نام و نام خانوادگی	محمد جواد رنجبر
شماره دانشجویی	۸۱۰۱۰۱۱۷۳
تاریخ ارسال گزارش	۱۴۰۲/۰۳/۱۵

فهرست گزارش

فهرست شکل ها

تمام فایل‌های خواسته شده شامل CSVها در فولدر مربوط قرار داده شده است.

سوال ۱

پیش پردازش دادگان با استفاده ازpreprocess-fariseq

در این مرحله این کارها توسط fariseq انجام می‌شود:

توکن‌بندی: Fairseq عمل توکن‌بندی را انجام می‌دهد که در آن متن به واحدهای جداگانه‌ای مانند کلمات، زیرکلمات یا حروف تقسیم می‌شود.

ایجاد واژگان: Fairseq واژگان بر اساس داده‌های توکن‌بندی شده ایجاد می‌کند. در این تمرین ۴۰۰۰۰ واژه تولید می‌کنیم.

عددسازی: بعد از توکن‌بندی، Fairseq یک نمایش عددی به هر توکن اختصاص می‌دهد.

آموزش مدل با استفاده از train-fairseq

پارامترهای داده شده در سوال را تنظیم می‌کنیم:

معماری یک لایه encoder و decoder و یک لایه lstm با مکانیزم attention

تابع هزینه label smooth cross entropy به مقدار label smoothing برار ۰.۲

نرخ یادگیری تطبیقی inverse sqrt با مقدار اولیه 0.0025

بهینه‌ساز adam با پارامترهای beta1=0.9 و beta2=0.98

Dropout با مقدار ۰.۲۵

برای ۶ epoch آموزش گذاشتیم.

استفاده از generate-fairseq برای ترجمه دادگان ارزیابی

تعدادی نمونه از این ترجمه‌ها به صورت زیر می‌باشد:

S-7043 it is not bad either ,

T-7043 ,این هم بد نیست

H-7043 -1.0841904878616333 ,این بد نیست

D-7043 -1.0841904878616333 ,این بد نیست

P-7043 -2.7383 -1.1646 -0.4039 -0.7685 -0.3457

S-7021 yes , Friday the fifth .

T-7021 . بله , جمعه پنجم

H-7021 -0.5548604726791382 . بله , جمعه پنجم

D-7021 -0.5548604726791382 . بله , جمعه پنجم

P-7021 -0.1186 -0.1450 -1.9711 -0.2583 -0.4583 -0.3779

S-6924 yes , that is okay .

T-6924 . بله , خوب است

H-6924 -0.5492016077041626 . بله , خوب است

D-6924 -0.5492016077041626 . بله , خوب است

P-6924 -0.1230 -0.1799 -2.3391 -0.1724 -0.2090 -0.2718

S-6920 which Friday do you mean ?

T-6920 ? منظورتان کدام جمعه است

H-6920 -1.489291787147522 ? که جمعه به معنای آن است

D-6920 -1.489291787147522 ? که جمعه به معنای آن است

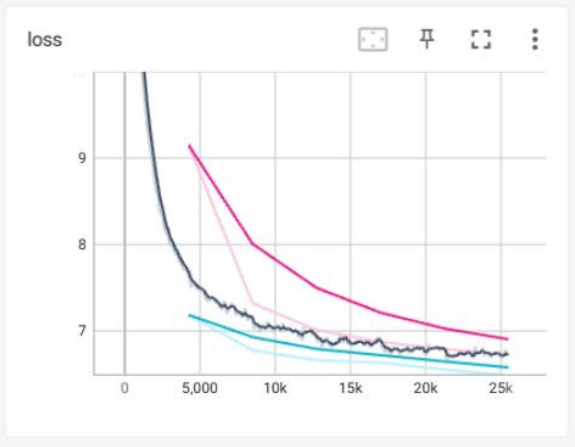
P-6920 -2.0477 -0.1897 -3.9687 -1.8533 -2.1071 -0.8938 -0.6505 -0.2035

S-6911 yes , at what time ?

T-6911 ? بله , چه ساعتی

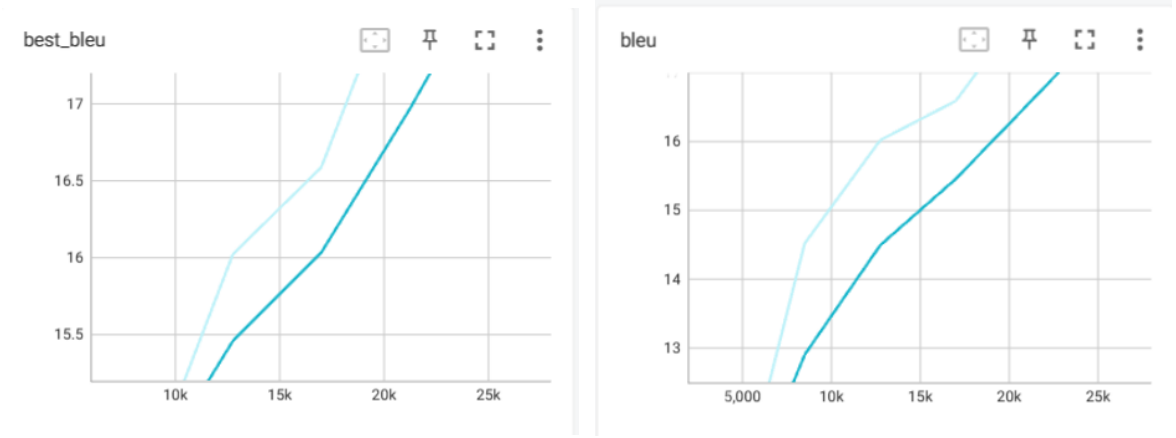
استفاده از Tensorboard برای بررسی فرایند آموزش

loss به شکل زیر است:



شکل ۱ **loss** بدون استفاده از **bpe**

نمودار BLEU برای دادگان ارزیابی به شکل‌های زیر می‌باشد:



شکل ۲ **BLEU** بدون **BPE**

همچنین مقدار نهایی این تابع برای داده‌های validation به شکل زیر است.

Generate test with beam=5: BLEU4 = 24.73, 58.5/31.7/18.3/11.0 (BP=1.000, ratio=1.093, syslen=99070, reflen=90650)

بر اساس جدول زیر می‌توان راجع به گفت مدل تقریبا مناسب کار می‌کند.

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors

BLEU Score	Interpretation
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

آموزش مدل BPE برای دادگان انگلیسی و فارسی

ابتدا با استفاده از SentencePieceTrainer کلمات را با روش BPE کلات را توکنایز می‌کنیم.

پردازش دادگان با مدل هایBPE

حال همانند بخش قبل داده‌ها را preprocess می‌کنیم که کارهایی که انجام می‌شد، ذکر شده است.

آموزش مدل با استفاده ازtrain-fairseq

پارامترهای داده شده در سوال را تنظیم می‌کنیم:

معماری یک لایه encoder و decoder و یک لایه lstm با مکانیزم attention

تابع هزینه label smooth cross entropy به مقدار label smoothing برار ۰.۲

نرخ یادگیری تطبیقی inverse sqrt با مقدار اولیه 0.0025

بهینه‌ساز adam با پارامترهای beta1=0.9 و beta2=0.98

Dropout با مقدار ۰.۲۵

برای ۶ epoch آموزش گذاشتیم.

استفاده از generate-fairseq برای ترجمه دادگان ارزیابی

نمونه‌ای از جملات ترجمه شده به صورت زیر می‌باشد.

S-3414 ▁that ▁would ▁be ▁fine.▁

T-3414 ▁ خوب ▁ است ▁.

H-3414 -۰.۷۵۹۴۴۹۰۰۵۱۲۶۹۵۳۱ ▁ که ▁ خوب ▁ است ▁.

D-3414 -۰.۷۵۹۴۴۹۰۰۵۱۲۶۹۵۳۱ ▁ که ▁ خوب ▁ است ▁.

P-3414 ۰.۲۵۹۴- ۰.۴۱۲۳- ۱.۰۹۰۴- ۰.۳۵۸۶- ۰.۵۵۱۷- ۰.۲۸۴۴- ۲.۸۵۸۹- ۰.۲۵۹۹-

S-3406 _do _you _know _that?_

T-3406 >> _ آنرا _unk?_ <<

H-3406 ۰.۸۰۳۱۸۴۷۴۷۶۹۵۹۲۲۹- ?_ آیا _ می _ دانید _?

D-3406 ۰.۸۰۳۱۸۴۷۴۷۶۹۵۹۲۲۹- ?_ آیا _ می _ دانید _?

P-3406 ۰.۲۲۰۱- ۱.۲۷۶۲- ۰.۵۴۶۹- ۰.۳۴۱۴- ۰.۹۳۴۳- ۰.۳۳۹۵- ۲.۴۲۴۷- ۰.۳۴۲۳-

S-2996 _at _two _P M._

T-2996 ._ دو _ بعدازظهر _

H-2996 ۰.۸۰۱۹۴۶۹۳۸۰۳۷۸۷۲۳- ._ در _ دو _ بعدازظهر _

D-2996 ۰.۸۰۱۹۴۶۹۳۸۰۳۷۸۷۲۳- ._ در _ دو _ بعدازظهر _

P-2996 ۰.۳۰۰۳- ۰.۷۶۵۱- ۲.۴۶۸۴- ۰.۳۶۳۶- ۰.۳۱۵۰- ۰.۳۰۳۴- ۱.۶۳۰۵- ۰.۲۶۹۳-

S-2710 _that _would _be _great._

T-2710 ._ عالی _ است _

H-2710 ۰.۹۵۰۵۳۹۹۴۶۵۵۶۰۹۱۳- ._ این _ عالی _ است _

D-2710 ۰.۹۵۰۵۳۹۹۴۶۵۵۶۰۹۱۳- ._ این _ عالی _ است _

P-2710 ۰.۲۳۸۷- ۰.۴۱۶۹- ۱.۲۸۴۲- ۰.۳۳۷۲- ۱.۱۷۸۹- ۰.۳۱۱۰- ۳.۵۷۵۱- ۰.۲۶۲۴-

S-2660 _you _will _find _that._

T-2660 ._ پیدا _ میکنید _

H-2660 ۰.۸۰۹۲۳۱۹۹۶۵۳۶۲۵۴۹- ._ شما _ پیدا _ خواهید _ کرد _

D-2660 ۰.۸۰۹۲۳۱۹۹۶۵۳۶۲۵۴۹- ._ شما _ پیدا _ خواهید _ کرد _

P-2660 - ۰.۱۷۶۵- ۰.۲۸۹۵- ۰.۹۹۱۸- ۰.۲۴۶۵- ۳.۷۴۷۴- ۰.۲۵۵۴- ۱.۳۸۴۱- ۰.۲۹۲۴-
۰.۲۴۸۳- ۰.۴۶۰۵

S-2676 _the _end _of _May._

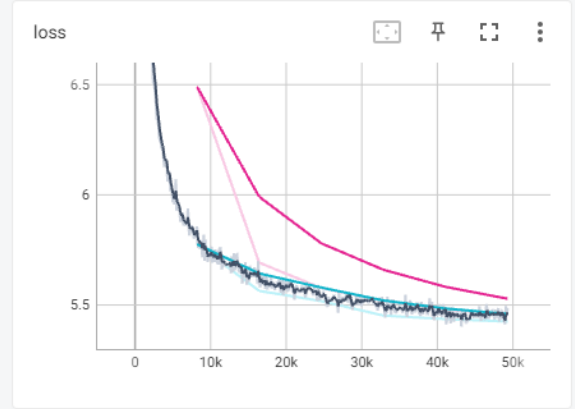
T-2676 ـــ اواخر ـــ می ـــ.

H-2676 -۰.۵۳۱۲۶۱۸۶۱۳۲۴۳۱۰۳ ـــ پایان ـــ ماه ـــ می ـــ.

D-2676 -۰.۵۳۱۲۶۱۸۶۱۳۲۴۳۱۰۳ ـــ پایان ـــ ماه ـــ می ـــ.

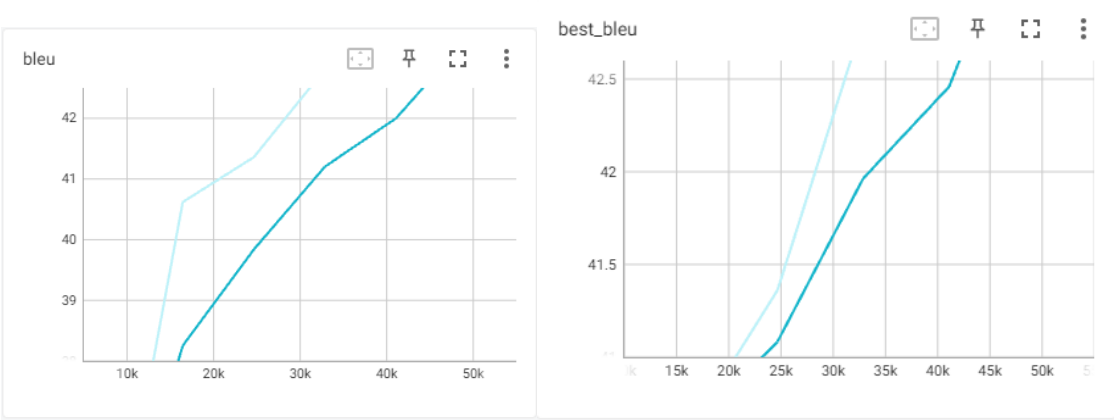
استفاده از Tensorboard برای بررسی فرایند آموزش

Loss به شکل زیر می‌باشد:



شکل ۳ **loss** مدل با توکنایزشن **BPE**

همچنین نمودار BLEU به شکل زیر خواهد بود:



شکل ۴ **BLEU** با توکنایزشن **BLEU**

همچنین مقدار نهایی BLEU برای داده‌های ارزیابی به صورت زیر می‌باشد:

Generate test with beam=5: BLEU4 = 42.91, 69.4/49.7/37.8/26.0 (BP=1.000, ratio=1.093, syslen=186038, reflen=170167)

این مقدار نشان می‌دهد که مدل به شدت خوب عمل می‌کند و BPE روش بهتری بود.

سوال ۲

دانلود Bert Model و Bert Tokenizer

ابتدا با استفاده از کتابخانه transformers مدل Bert multilingual را لود می‌کنیم.

پردازش دادگان AFEC با استفاده از Bert Tokenizer

حال داده‌ها را به ترتیب با استفاده از این مدل tokenize می‌کنیم و در پوشه‌ای جدید با نام Tokenized_data ذخیره می‌کنیم.

پیش پردازش دادگان با استفاده از preprocess-fariseq

مانند سوال قبل نیز در این مرحله مراحل پیش‌پردازش توسط fariseq انجام می‌شود.

ذخیره وزن های اولیه embedding شبکه Bert با فرمت مناسب

از آنجا که تعداد کلمات مدل bert خیلی زیاد بود و colab با این تعداد embedding را نمی‌توانست در حافظه نگه دارد، فقط ۵۰۰۰ تا از کلمه‌ها را نگه می‌داریم. و بر اساس فرمت خواسته شده در سوال این کلمات را ذخیره می‌کنیم.

آموزش مدل با استفاده از fairseq-train

ابتدا با استفاده از دستورات زیر مدل را برای آموزش با embedding استخراج شده از Bert را لود کرده و بر اساس آن‌ها آموزش می‌دهیم. در این مرحله یک بار مدل را freeze کردن و یک بار بدون این کار آموزش می‌دهیم،

Freeze شده

در این نحوه با استفاده از دستورات زیر بخش‌هایی از مدل را فریز می‌کنیم.

```
    --encoder-freeze-embed \
    --decoder-freeze-embed \
```

Freeze نشده

استفاده از generate-fairseq برای ترجمه دادگان ارزیابی

تعدادی از نتایج به صورت زیر می‌باشد.

Freeze شده

S-6392 that is oka ##y.

T-6392 . خوب است

H-6392 ۱.۹۰۷۴۸۷۳۹۲۴۲۵۵۳۷- آن بسیار خوب است .

D-6392 ۱.۹۰۷۴۸۷۳۹۲۴۲۵۵۳۷- آن بسیار خوب است .

P-6392 ۰.۳۱۳۱- ۰.۲۰۶۵- ۰.۱۱۹۱- ۲.۴۶۵۲- ۳.۱۵۱۳- ۵.۱۸۹۷-

S-6370 that is oka ##y.

T-6370 . خوب است

H-6370 ۱.۹۰۷۴۸۷۳۹۲۴۲۵۵۳۷- آن بسیار خوب است .

D-6370 ۱.۹۰۷۴۸۷۳۹۲۴۲۵۵۳۷- آن بسیار خوب است .

P-6370 ۰.۳۱۳۱- ۰.۲۰۶۵- ۰.۱۱۹۱- ۲.۴۶۵۲- ۳.۱۵۱۳- ۵.۱۸۹۷-

S-6321 is that oka ##y?

T-6321 ? خوب است

H-6321 ۱.۸۴۸۶۶۴۷۶۰۵۸۹۵۹۹۶- این م ##سا ##له است ?

D-6321 ۱.۸۴۸۶۶۴۷۶۰۵۸۹۵۹۹۶- این م ##سا ##له است ?

P-6321 ۰.۲۰۴۴- ۰.۲۴۱۲- ۰.۷۱۹۷- ۰.۱۷۸۷- ۲.۴۰۸۳- ۶.۴۲۳۴- ۲.۷۶۴۸-

S-5998 that is oka ##y.

T-5998 . خوب است

H-5998 ۱.۹۰۷۴۸۷۳۹۲۴۲۵۵۳۷- آن بسیار خوب است .

D-5998 ۱.۹۰۷۴۸۷۳۹۲۴۲۵۵۳۷- آن بسیار خوب است .

P-5998 ۰.۳۱۳۱- ۰.۲۰۶۵- ۰.۱۱۹۱- ۲.۴۶۵۲- ۳.۱۵۱۳- ۵.۱۸۹۷-

S-8503 make a suggest ##ion.

T-8503 . پیشنهاد ب ##ده

Freeze نشده

P-7234 ۰.۱۷۵۷- ۰.۲۲۳۳- ۰.۳۸۲۴- ۱.۸۶۲۰- ۰.۹۴۶۲- ۰.۰۸۸۰- ۱.۶۹۸۷- ۰.۳۸۷۹- ۰.۲۷۵۷

S-7012 in the afternoon it is still possible.

T-7012 . بعد از ظهر میشود

H-7012 ۰.۸۳۸۹۸۸۹۵۹۷۸۹۲۷۶۱- . در بعد ##از ##ظهر هنوز امکان ##پ ##ذیر است

D-7012 ۰.۸۳۸۹۸۸۹۵۹۷۸۹۲۷۶۱- . در بعد ##از ##ظهر هنوز امکان ##پ ##ذیر است

P-7012 ۲.۳۰۹۲- ۰.۵۲۷۸- ۱.۵۷۹۲- ۰.۳۶۶۸- ۱.۸۹۳۹- ۰.۹۶۹۷- ۰.۴۴۱۹- ۰.۱۱۹۵- ۰.۳۶۱۴ ۰.۳۸۱۶- ۰.۲۷۸۰-

S-6559 ye ##s , that would be possible.

T-6559 . بل ##ه , میشود

H-6559 ۰.۶۷۱۱۰۷۳۵۱۷۷۹۹۳۷۷- . بل ##ه , امکان ##پ ##ذیر است

D-6559 ۰.۶۷۱۱۰۷۳۵۱۷۷۹۹۳۷۷- . بل ##ه , امکان ##پ ##ذیر است

P-6559 ۰.۱۷۵۷- ۰.۲۲۳۳- ۰.۳۸۲۴- ۱.۸۶۲۰- ۰.۹۴۶۲- ۰.۰۸۸۰- ۱.۶۹۸۷- ۰.۳۸۷۹- ۰.۲۷۵۷

S-6229 that is a very good suggest ##ion.

T-6229 . پیشنهاد بسیار خوبی است

H-6229 ۱.۰۶۲۴۰۹۲۸۱۷۳۰۶۵۱۹- . این یک پیشنهاد بسیار خوب است

D-6229 ۱.۰۶۲۴۰۹۲۸۱۷۳۰۶۵۱۹- . این یک پیشنهاد بسیار خوب است

P-6229 ۲.۵۲۱۹- ۱.۶۳۶۸- ۱.۱۰۳۹- ۱.۲۷۷۳- ۰.۵۶۴۷- ۰.۵۷۹۹- ۰.۴۹۶۹- ۰.۳۱۷۹-

S-2330 ten o ' clock in the morning.

T-2330 ساعت ده ص ##بج .

H-2330 -۰.۴۷۰۹۹۸۰۴۸۷۸۲۳۴۸۶۳ ساعت ده ص ##بج .

D-2330 -۰.۴۷۰۹۹۸۰۴۸۷۸۲۳۴۸۶۳ ساعت ده ص ##بج .

P-2330 -۰.۰۶۵۱ -۰.۳۸۰۲ -۰.۰۶۰۴ -۰.۰۵۴۹ -۱.۸۰۱۴ -۰.۴۶۴۰

S-1559 this is on the fifth of January.

T-1559 میشود پنج ##م ژانویه .

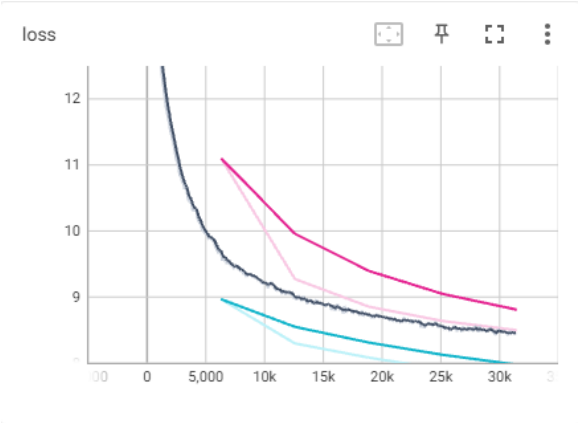
H-1559 -۰.۷۹۰۴۷۲۶۸۶۲۹۰۷۴۱ این پنج ##م ژانویه است .

D-1559 -۰.۷۹۰۴۷۲۶۸۶۲۹۰۷۴۱ این پنج ##م ژانویه است .

استفاده از Tensorboard برای بررسی فرایند آموزش

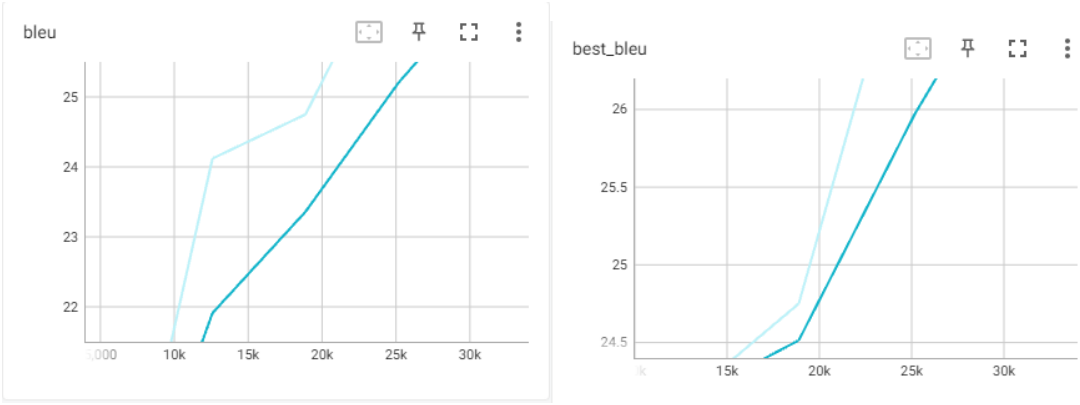
Freeze شده

Loss به شکل زیر می‌باشد:



شکل ۵ **loss** مدل **freeze** شده و استفاده از **Bert embedding**

نمودار BLEU برای دادگان ارزیابی به شکل‌های زیر می‌باشد:



شکل ۶ **Bleu** مدل **freeze** شده و استفاده از **Bert embedding**

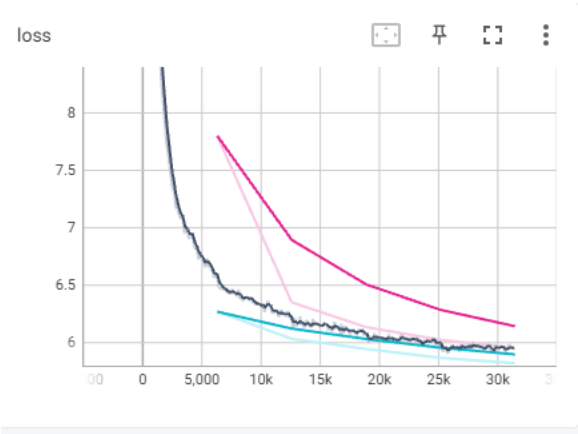
همچنین مقدار نهایی Bleu برای داده‌های ارزیابی به صورت زیر می‌باشد:

Generate test with beam=5: BLEU4 = 17.46, 42.8/22.6/13.1/7.3 (BP=1.000, ratio=1.140, syslen=156777, reflen=137543)

که مدل آنچنان خوب نیست.

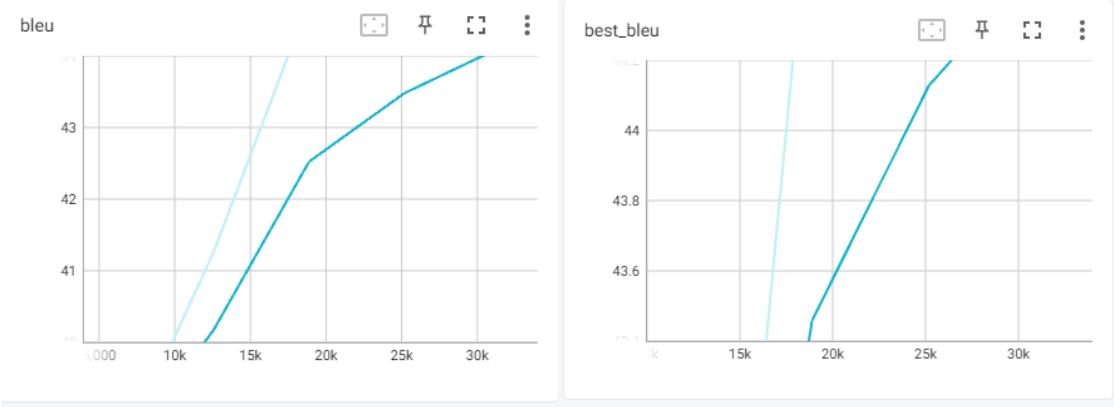
Freeze نشده

Loss به شکل زیر می‌باشد:



شکل ۷ **loss** مدل **freeze** نشده و استفاده از **Bert embedding**

نمودار BLEU برای دادگان ارزیابی به شکل‌های زیر می‌باشد:



شکل ۸ **Bleu** مدل **freeze** نشده و استفاده از **Bert embedding**

مقدار نهایی Bleu برای داده‌های ارزیابی به صورت زیر می‌باشد:

Generate test with beam=5: BLEU4 = 35.53, 62.9/41.9/29.5/20.5 (BP=1.000, ratio=1.012, syslen=139244, reflen=137543)

که با توجه به این نمره مشخص است که به شدت از حالت قبلی مدل بهتر و ترجمه‌های قابل قبولتری دارد و با توجه به فریز نشدن مدل این انتظار می‌رفت.