به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس پردازش متن و زبان طبیعی

سحر رجبی
شماره دانشجویی
۸۱۰۱۹۹۱۶۵

خرداد ماه 1400

شرح پروژه

مدل‌های انتخابی:

برای انجام پروژه، ما از دو ابزار OpenNMT و FairSeq استفاده کردیم (البته این پروژه به صورت تک نفره انجام شده، و طبق گفته‌ی استاد درس، در صورت انجام انفرادی، بررسی یک ابزار کافی است). توضیحات مربوط به هر یک از ابزارها در ذیل بخش‌های خواسته‌شده نوشته‌شده است.

مراحل پیاده‌سازی با ابزار OpenNMT:

مرحله‌ی اول بعد از نصب OpenNMT و sentencepiece، آماده‌سازی داده‌ها برای آموزش مدل است. برای این کار، یک اسکریپ (با نام prepare_data.sh) نوشته‌شده که با استفاده از ابزار sentencepiece در سطح subword و unigram مدلی می‌سازیم که برای ساخت vocab از آن استفاده کنیم. دلایل استفاده از sentencepiece را در ادامه بررسی خواهیم کرد. بعد از ساخت این مدل، مجموعه لغات دو زبان مبدا و مقصد به کمک آن به‌دست‌می‌آیند. سپس با استفاده از کانفیگی که در فایل config.yaml قابل مطالعه است، با مشخص کردن پارامترهای مختلف، مدل خود را آموزش می‌دهیم.

در نهایت برای تست مدل، باید داده‌های تست را هم مطابق روشی که داده‌های آموزش و ولیدیشن tokenize شدند، پراسس کنیم و نهایتا از آن برای ارزیابی نتیجه‌ی مدل، استفاده کنیم.

مراحل پیاده‌سازی با ابزار FairSeq:

برای کار با این ابزار، می‌توانیم از کامند fairseq-preprocess استفاده کنیم، که با دریافت کردن دادگان و زبان مبدا و مقصد، می‌تواند مدل لازم برای tokenize کردن را ایجاد کند. در اینجا با استفاده از پارامتر bpe– از sentencepiece استفاده کردیم. سپس با استفاده از کامند fairseq-train مدل خود را آموزش دادیم.

تست ابزارها:

مراحل انجام‌شده برای تست ابزارها، در داخل فایل test.sh در فولدر مربوط به هر کدام از آن‌ها موجود است که در نهایت معیار Bleu را محاسبه می‌کند.

سوال ۱:

آخرین نسخه از مدل آموزش دیده توسط OpenNMT دارای معیار Bleu برابر با ۰.۹ است و خروجی آن بر روی دادگان تست به صورت کلی به این صورت است:

```
BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1 = 0.9 20.5/2.1/0.3/0.0 (BP = 1.000 ratio = 1.007 hyp_len = 11470 ref_len = 1138
```

و نمونه‌ای از جملات ترجمه شده توسط این مدل:

yingluck won 296 votes in the nearly 500 - member parliament.

وی در سال 1915 میلیون دلاری در مجلس سنا رای دهند.

more than 90 people were killed and hundreds were injured.

پنج نفر کشته شدند و بیش از 80 نفر زخمی شدند.

از طرف دیگر، آخرین مدل آموزش دیده با ابزار FairSeq دارای معیار Bleuی برابر ۱.۸۵ است و خروجی آن بر روی دادگان تست به این صورت است:

```
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint100.pt
Generate test with beam=5: BLEU4 = 1.85, 27.0/3.6/0.9/0.2 (BP=0.902, ratio=0.906, syslen=10319, reflen=11385)
```

نمونه‌ای از جملات ترجمه شده توسط این ابزار در زیر آورده‌شده است:

more than 90 people were killed and hundreds were injured .

بیش از دوازده نفر کشته شدند و صدها نفر در بازداشت شدند.

yingluck won 296 votes in the nearly 500 - member parliament

مدت کوتاهی پس از اعلام اظهارات wmc وابسته به عضو کابینه عضو سازمان ملل منتشر کرده‌است.

مشاهده‌ی دو مثال برای مقایسه‌ی دو مدل کافی نیست؛ اما ترجمه‌ی این دو جمله توسط مدل OpenNMT به مراتب بهتر بوده و معیار Bleu مقدار بیشتری در FairSeq دارد.

یکی از علل این تفاوت، می‌تواند نیاز به ایپاک‌های بیشتر باشد. و یا معماری استفاده‌شده می‌تواند تاثیرگذار باشد. در ضمن ابزار FairSeq تعداد قابل توجهی از کلمات را به عنوان <unk> می‌شناسد. در جمله‌ی دوم، کلمه‌ی Yingluck و 296 هر دو به این صورت هستند و این خود می‌تواند تاثیر در خروجی داشته‌باشد.

سوال ۲:

در وهله‌ی اول، از نظر سرعت آموزش مدل، ابزار FairSeq می‌توانست با روش‌هایی مثل استفاده از fp16 که بر روی gpu قابل استفاده است و یا نصب Nvidia Apex سرعت را تا حد خیلی زیادی افزایش دهد و سرعت یادگیری در آن نسبت به OpenNMT بیشتر است.

هر دوی این ابزارها اجازه‌ی اعمال تغییرات تا حد نیاز ما -و فراتر- را می‌دهند و در مقایسه‌ی documentationها هم هر دو خوب عمل کرده‌اند. اما استفاده از FairSeq پیچیدگی کمتری داشت.

سوال ۳:

- در ابزار OpenNMT پارامترهای موثر در training که tune شده‌اند، به شرح زیر هستند:

- Src_subword_nbest و tgt_subword_nbest که تعیین می‌کند در هنگام استفاده از مدل subwordهای آموزش داده‌شده، چندتا از بهترین کاندیدها در هر یک از دو زبان مبدا و مقصد انتخاب شوند.
- Src_subword_alpha و tgt_subword_alpha که درجه‌ی smoothing در مدل unigram در subwordها را مشخص می‌کند.
- Src_seq_length و tgt_seq_length ماکسیمم طول جملات در هر یک از زبان‌های مبدا و مقصد را مشخص می‌کند.
- Save_checkpoint_steps مشخص می‌کند که بعد از هر چند step یک checkpoint از مدل ذخیره شود تا در ادامه بتوان از آن استفاده کرد.

- Train_steps در واقع تعداد iterationها را تعیین می‌کند (با epochs متفاوت است چرا که هر batch یک iteration است).
- Valid_steps مشخص می‌کند که بعد از چند iteration نیاز است که مدل را روی دادگان ارزیابی تست کنیم و عملکرد آن را گزارش کنیم.
- Warmup_steps تعداد iterationها قبل از شروع کاهش نرخ یادگیری است.
- Decoder_type، encoder_type، word_vec_size، rnn_size، layers و transformer_ff پارامترهای تعیین معماری مدل هستند.
- Accum_count تعیین می‌کند که gradient update بعد از چند batch صورت بگیرد.
- Optim الگوریتم مورد استفاده برای بهینه‌سازی را تعیین می‌کند و پارامترهای adam_bata1 و adam_beta2 مربوط به آن هستند.
- Decay_method روش کاهش نرخ یادگیری را مشخص می‌کند و learning_rate هم خود نرخ یادگیری است.
- Batch_size اندازه‌ی batch را تعیین می‌کند.
- Dropout احتمال رخداد droput در لایه‌ها را تعیین می‌کند.

• از بین این متغیرها، مقادیری که با مقدار دیفالت متفاوت هستند به این شرح است:

- Src_subword_nbest و tgt_subword_nbest هر کدام برابر ۳ هستند.
- Src_seq_length و tgt_ seq_length هرکدام برابر ۱۵۰ هستند.
- مقدار save_checkpoints_step برابر ۱۰۰۰ ایتریشن است.
- Train_steps با توجه به محدودیت منابع، برابر با ۱۵۰۰۰ در نظر گرفته شده است.
- Valid_steps برابر با ۱۰۰۰ است که یعنی بعد از هر ۱۰۰۰ step، مدل بر روی دادگان ارزیابی اعمال خواهد شد.
- Warmup_steps هم در اینجا برابر با ۱۰۰۰ قرار داده‌شده است.
- Accum_count برابر با ۴ قرار داده شده‌است.
- برای optim از روش adam استفاده کرده‌ایم.
- مقدار dropout را برابر با ۰.۲ تنظیم کردیم.

• در ابزار FairSeq پارامترهای موثر در training که tune شده‌اند، به شرح زیر هستند:

- Lr یا همان نرخ‌یادگیری که اندازه‌ی گام‌ها برای به‌روزرسانی وزن‌ها بعد از محاسبه‌ی گردایان را مشخص می‌کند.
- Clip-norm کنترل می‌کند که نرم l2 گرادیان از مقدار مشخصی کمتر نشود تا شاهد vanishing gradient نباشیم.

- Dropout احتمال اعمال dropout در لایه‌ها را مشخص می‌کند.
- Max_tokens تعیین می‌کند که بیشینه تعداد token در یک batch چه مقداری باشد. این پارامتر در زمانی که gpu محدودی داریم، می‌تواند اثرگذار باشد.
- Arch که معماری انتخابی برای آموزش مدل را مشخص می‌کند.
- Optimizer روش بهینه‌سازی در مراحل آموزش را تعیین می‌کند.
- Bpe ابزار مورد استفاده برای byte-pair encoding را مشخص می‌کند.
- Max-epoch بیانگر بیشنیه تعداد مجاز epoch در روند آموزش است.
- Save_interval بیان می‌کند که بعد از چند epoch یک checkpoint از مدل را ذخیره کنیم.
- Fp16 با استفاده از halp precision floating point میتواند محاسبات را بهینه‌تر بر روی gpu انجام بدهد.

• از بین این متغیرها، مقادیری که با مقدار دیفالت متفاوت هستند به این شرح است:

- مقدار lr برابر با ۰.۰۰۰۱ است.
- Clip-norm برابر ۰.۱ تعیین شده است.
- پارامتر dropout را برابر با ۰.۲ در نظر گرفتیم.
- Max_tokens را به علت حافظه‌ی محدود، تنها تا مقدار ۸۰۰۰ بالا بردیم.
- Arch برابر transformer ست شده است.
- Optimizer استفاده شده در این ابزار adam است.
- برای bpe از ابزار sentencepiece استفاده کرده‌ایم.
- به علت محدودیت زمان و فضای ذخیره‌سازی checkpointها، مقدارmax_epochs را برابر با ۱۰۰ قرار دادیم.
- بعد از هر ۱۰ ایپاک (مقدار save_interval) یک checkpoint از مدل را ذخیره می‌کنیم.
- همچنین از قابلیت fp16 برای بالا بردن سرعت محاسبات استفاده کرده‌ایم.

سوال ۴:

در مدل آموزش دیده توسط ابزار OpenNMT، از نظر من پارامترهای زیر می‌تواند تاثیر مستقیم بر خروجی مدل ما داشته‌باشد:

1 - مقدار dropout اهمیت بالایی دارد؛ چرا که می‌تواند از overfit شدن مدل ما جلوگیری کند (در واقع این پارامتر احتمال حذف یک نورون از معماری را تعیین می‌کند. با حذف رندم این بخش‌ها، می‌توانیم از overfitting جلوگیری کنیم)

2 - پارامتر src_subword_nbest و tgt_subword_nbest از آنجایی که نقش مستقیم در tokenization ما دارند؛ تعیین آن‌ها می‌تواند نوع برخور مدل با داده‌ها را دست‌خوش تغییر کند و نتیجه‌ی نهایی را متفاوت کند (توضیح آن‌ها در قسمت قبل آورده‌شده‌است).

3 - الگوریتم استفاده‌شده برای بهینه‌سازی (optim) تاثیر مستقیم بر روند به‌روزرسانی وزن‌ها و در نتیجه، خروجی مدل خواهد داشت. که در اینجا ما از روش adam استفاده کرده‌ایم.

4 - Decay method که به تدریج learning rate ما را با الگوریتم مشخص‌شده کاهش می‌دهد؛ باعث کنترل اندازه‌ی گام‌ها در بروزرسانی وزن‌ها شده و با کاهش آن، احتمال اینکه ما از بهینه‌ی تابع عبور کنیم و به آن نزدیک نشویم را کاهش می‌دهد.

5 - Batch-size پارامتر مهمی است که بین سرعت محاسبات و convergence یک تعادل ایجاد می‌کند و تعیین مناسب آن برای رسیدن به سرعت مناسب و احتمال convergence بالا اهمیت به‌سزایی دارد.

در مدل آموزش دیده توسط ابزار FairSeq، از نظر من پارامترهای زیر می‌تواند تاثیر مستقیم بر خروجی مدل ما داشته‌باشد:

1 - مقدار learning rate از آنجایی که اندازه‌ی گام‌های بروزرسانی برای وزن‌ها را تعیین می‌کند، تاثیر ویژه و مهمی دارد. چرا که در عین حال که مقدار بالا می‌تواند باعث افزایش سرعت بروزرسانی شود، در عین حال می‌تواند باعث رد کردن نقاط می‌نیمم و همگرا نشدن مدل باشد.

2 - مقدار dropout مشابه توضیحات قسمت قبل، از آنجایی که میتواند جلوی overfitting در مدل را بگیرد اهمیت زیادی دارد.

3 - پارامتر max_tokens که در واقع چیزی شبیه به batch_size است مشابه توضیحات قسمت قبل می‌تواند سرعت آموزش و همگرایی را تحت تاثیر بگذارد و از این نظر پارامتر مهمی است.

4 - الگوریتم optimization که در اینجا هم روش adam است به طور مستقیم در روند به‌روزرسانی وزن‌ها اثر دارد و خروجی مدل به آن وابسته است.

5 - مقدار clip-norm هم از آنجایی که می‌تواند مانع vanishing gradient شود روند آموزش را تحت تاثیر می‌گذارد و خروجی مدل میتواند متاثر از آن باشد.

سوال ۵:

برای ترجمه‌ی جملات، توانایی فهم کلمات دیده نشده اهمیت بالایی دارد. یکی از روش‌هایی که می‌توان داده‌ها را با توجه به آن به توکن تبدیل کرد، روش‌های byte-pair encoding هستند. این روش‌ها با توجه به فرکانس حروف دیده شده، سعی می‌کنند subwordهایی که احتمالا در ساخت کلمات استفاده می‌شود را تشخیص دهد.

در این تمرین، ما با استفاده از ابزار sentencepiece دو مدل از روی دادگان آموزش ساختیم که subwordها را برای هر دو زبان مبدا و مقصد تشخیص دهد (این مدل برای هر دو زبان فارسی و

انگلیسی مناسب است. چرا که وابسته به ساختار زبان نیست و از روی ظاهر جملات، زیرکلمات را بدست می‌آورد) برای مثال، در زبان فارسی شناسه‌ی افعال و یا در زبان انگلیسی ing و .. می‌توانند به عنوان subwordها در فهم کلمات جدید کمک کنند.

از آنجایی که برای هر دوی ابزارها ، از ابزار sentencepiece استفاده شده؛ یک نسخه از توکن‌های تشخیص داده‌شده بعد از این پیش‌پردازش‌ها، در فایل ارسالی ضمیمه شده است.

سوال ۶:

در تصویر زیر، می‌توان روند تغییرات معیار Bleu را در ابزار FairSeq بررسی کرد.

```
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint10.pt
Generate test with beam=5: BLEU4 = 0.00, 21.7/2.4/0.3/0.0 (BP=0.878, ratio=0.885, syslen=10079, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint20.pt
Generate test with beam=5: BLEU4 = 1.20, 25.7/3.3/0.6/0.1 (BP=0.818, ratio=0.833, syslen=9481, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint30.pt
Generate test with beam=5: BLEU4 = 1.31, 25.4/3.4/0.6/0.1 (BP=0.955, ratio=0.956, syslen=10889, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint40.pt
Generate test with beam=5: BLEU4 = 1.56, 27.3/3.5/0.7/0.1 (BP=0.892, ratio=0.897, syslen=10216, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint50.pt
Generate test with beam=5: BLEU4 = 1.58, 24.5/3.1/0.6/0.1 (BP=1.000, ratio=1.089, syslen=12393, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint60.pt
Generate test with beam=5: BLEU4 = 1.25, 26.3/3.2/0.6/0.1 (BP=0.919, ratio=0.922, syslen=10493, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint70.pt
Generate test with beam=5: BLEU4 = 1.61, 26.2/3.3/0.7/0.1 (BP=0.965, ratio=0.965, syslen=10990, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint80.pt
Generate test with beam=5: BLEU4 = 1.78, 27.9/3.7/0.9/0.2 (BP=0.837, ratio=0.849, syslen=9665, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint90.pt
Generate test with beam=5: BLEU4 = 1.87, 27.2/3.4/0.8/0.2 (BP=0.906, ratio=0.910, syslen=10360, reflen=11385)
# Translating with checkpoint /content/drive/MyDrive/data/checkpoints/fconv/checkpoint_last.pt
Generate test with beam=5: BLEU4 = 1.85, 27.0/3.6/0.9/0.2 (BP=0.902, ratio=0.906, syslen=10319, reflen=11385)
```

همانطور که مشخص است، مقدار آن به تدریج افزایش پیدا کرده است و در نهایت به ۱.۸۵ رسیده است.