به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس پردازش متن و زبان طبیعی

تمرین سوم

مدل زبانی مبتنی بر LSTM

بهار 1400

0- فهرست

1- دادگان و پیش‌پردازش

برای این تمرین از دادگان موجود در این آدرس[1] استفاده خواهیم کرد. برای مطلع شدن از چیستی این دادگان، می‌توانید توضیحات لینک فوق را مطالعه کنید. محتوای مورد نیاز ما از طریق این لینک[2] قابل دانلود است. این دادگان، بدون برچسب یا گروه‌بندی، به صورت مجموعه متون خام ۱۷,۸۶۸ کتاب هستند. ما در این تمرین از کتاب "crocodiles-spirit" از این مجموعه استفاده می‌کنیم. ده درصد جملات را به عنوان دادگان ارزیابی و مابقی را به عنوان دادگان آموزشی در نظر خواهیم گرفت.

با توجه به اینکه در این تمرین، تمرکز بر پیش‌پردازش نیست، نسخه‌ی ساده‌ای از پیش‌پردازش در اختیار شما قرار خواهد گرفت. در این نوت‌بوک[3] کدهای مورد نیاز برای دانلود و پیش‌پردازش دادگان قرار دارد. می‌توانید این کد را متناسب با نیاز خود تغییر دهید.

[1] https://www.reddit.com/r/MachineLearning/comments/ji7y06/p_dataset_of_196640_books_in_plain_text_for/
[2] https://the-eye.eu/public/AI/pile_preliminary_components/books1.tar.gz
[3] https://colab.research.google.com/drive/1Qtq8UgUzYyJtqQL4iWksT4Hjm_e4bTPY?usp=sharing

2- ابزار مورد نیاز

در این تمرین، برای ساخت و استفاده شبکه‌های عصبی از pytorch استفاده خواهیم کرد. این framework برای شبکه‌های عصبی مورد استفاده در پردازش متن نیز انعطاف خوبی از خود نشان داده است. در ادامه لینک‌هایی برای مطالعه ابزار مورد نیاز قرار داده خواهد شد. بسته به دانش قبلی خود، ممکن است به برخی یا هیچکدام از این مطالب نیاز نداشته باشید:

1. مقدمه‌ای بر pytorch با تمرکز بر مباحث nlp[4]
2. یادگیری عمیق با pytorch با تمرکز بر مباحث nlp (طبقه‌بند Logistic Regression Bag-of-Words)[5]
3. جانمایی کلمات و مدل زبانی با pytorch (استفاده از word embeddings و ساخت Language Model)[6]
4. ماژول Embedding در pytorch[7]
5. ماژول LSTM در pytorch[8]
6. ماژول GRU در pytorch[9]
7. ماژول Glove (بردار کلمات از پیش آموزش داده شده)[10]؛ ما در این تمرین از نسخه با مشخصات `name=6B,dim=50` استفاده خواهیم کرد.
8. استفاده از GPU در pytorch[11]

مطالعه منابع 4، 5 و 6 در لیست بالا، برای این تمرین حتما پیشنهاد می‌شود. همچنین استفاده از GPU در این تمرین الزامیست. بنابراین پیشنهاد می‌شود پروژه خود را در محیط google colab انجام دهید تا از سرویس رایگان GPU بدون دشواری بهره‌مند شوید.

9. سرویس colab[12]
10. سرویس GPU در colab[13]

---

[4] https://pytorch.org/tutorials/beginner/nlp/pytorch_tutorial.html#sphx-glr-beginner-nlp-pytorch-tutorial-py
[5] https://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html#sphx-glr-beginner-nlp-deep-learning-tutorial-py
[6] https://pytorch.org/tutorials/beginner/nlp/word_embeddings_tutorial.html#sphx-glr-beginner-nlp-word-embeddings-tutorial-py
[7] https://pytorch.org/docs/stable/generated/torch.nn.Embedding.html
[8] https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html
[9] https://pytorch.org/docs/stable/generated/torch.nn.GRU.html
[10] https://pytorchnlp.readthedocs.io/en/latest/source/torchnlp.word_to_vector.html
[11] https://medium.com/ai%C2%B3-theory-practice-business/use-gpu-in-your-pytorch-code-676a67faed09
[12] https://colab.research.google.com/
[13] https://medium.com/deep-learning-turkey/google-colab-free-gpu-tutorial-e113627b9f5d

3- FFNNLM (اختیاری)

- این بخش حکم دست‌گرمی داشته و هنگام بررسی گزارش نادیده گرفته خواهد شد.

در این بخش، هدف پیاده‌سازی یک مدل زبانی مبتنی بر یک شبکه عصبی feed forward را داریم. برای ساخت این مدل زبانی، گام‌های زیر را طی کنید:

1. اندازه پنجره مشخص w (مثلا ۵) را در نظر بگیرید. از دادگان آموزشی، تمام w-gramهای ممکن را استخراج کنید.
2. از هر w-gram، قسمت‌های features,labels را بدست آورید.
3. کلمات قسمت features را به بردارهای one-hot تبدیل کنید.
4. یک شبکه عصبی با دو لایه‌ی خطی پنهان بسازید که اندازه‌ی خروجی لایه‌ی اول آن برابر ۵۰ باشد. شبکه عصبی شما باید در حکم یک طبقه‌بند عمل کند که کلاس‌های خروجی آن، لغات لغتنامه باشند.
5. شبکه را به کمک mini batch هایی حاوی w-gram ها آموزش دهید. با استفاده از cross entropy، مقدار Loss شبکه را محاسبه کرده و نمودار آن را حین آموزش رسم و کاهش آن را کنترل کنید.
6. مقدار Loss شبکه بر روی دادگان تست را بدست آورید و مقدار سرگشتگی[14] را محاسبه کنید.
7. گام 6 را حین آموزش تکرار و نمودار کاهش Loss به مرور آموزش را رسم کنید.
8. با استفاده از این شبکه، یک جمله جدید تولید[15] کنید. به نحوه شروع و پایان این فرایند دقت شود.

حال تغییرات زیر را در مدل اعمال کنید:

1. کل فرایند آموزش و ارزیابی مدل را به GPU منتقل و اختلاف سرعت اجرا را مشاهده کنید.
2. ماژول Embedding را جایگزین لایه‌ی خطی پنهان اول شبکه خود کرده و فرایند آموزش را طی کنید.
3. ماژول Embedding را از شبکه خود حذف و بجای آن از بردارهای تعبیه شده توسط ماژول GloVe استفاده کنید.

[14] Perplexity
[15] Generate

LSTM -4

مبنی بر دانش خود از شبکه‌های RNN و توسعه LSTM، لینک‌های ارائه شده در بخش دوم و ساختار کلی ارائه شده برای پیاده‌سازی مدل زبانی عصبی در بخش سوم، گام‌های مورد نیاز برای پیاده‌سازی مدل زبانی مبتنی بر LSTM را مشخص کرده و این مدل را پیاده‌سازی کنید. تأثیر حداقل چهار مورد از تغییرات زیر بر عملکرد مدل را بررسی کنید (تمام جوانب از جمله سرعت کاهش سرگشتگی، نقطه همگرایی، overfitting و توانایی مدل در تولید جمله را درنظر بگیرید؛ بررسی بیشتر از چهار مورد اختیاری و بدون امتیاز است):

1. اندازه لایه‌ی LSTM
2. تعداد لایه‌های LSTM
3. اندازه لایه‌ی Embedding
4. استفاده از Glove بجای آموزش Embedding
5. اندازه و تعداد لایه‌های خطی بعد از LSTM
6. اندازه و تغییرات learning rate
7. مقداردهی اولیه ماتریس‌های h و c (ماتریس‌های state در LSTM) در ابتدای یک جمله
8. استفاده از GRU بجای LSTM

به نکات زیر دقت کنید:

- مدل را بر روی GPU آموزش داده و ارزیابی کنید. سرعت عملکرد GPU چندین برابر CPU می‌باشد.
- حتما از mini batch training استفاده کنید.
- برای آموزش به صورت batch training، به راهکاری برای طول متفاوت جملات نیاز دارید. برای حل این معضل، راهکارهای متفاوتی وجود دارد. در صورتی که بیش از یک راه را امتحان کردید، مفید است اگر در مورد همه آن‌ها و تفاوتشان توضیح دهید.
- اگر از mini batch هایی با طول متفاوت استفاده می‌کنید، به نحوه محاسبه و اعمال loss در گام‌های متفاوت دقت کنید.
- انتخاب learning rate می‌تواند بسیار تاثیرگذار باشد.
- مهم است که هنگام مقایسه‌ی دو مدل که در یک نقطه با هم متفاوت‌اند، نمودارهای متناظر را در یک پلات رسم کنید (قرار دادن legend فراموش نشود).

- حدود مقادیر loss مشاهده شده به عواملی همچون پیش‌پردازش و یا طریقه حل معضل طول متفاوت جملات وابسته است (چرا؟). این موضوع را در فرایند مقایسه‌ی مدل‌ها فراموش نکنید.
- مواجهه با local minima بسیار محتمل است. بررسی این مسئله را فراموش نکنید. در صورت تشخیص local minima، خوب است اگر دلیل وجود آن را نیز به صورت شهودی توجیه کنید.
- توجه کنید که در این تمرین، تقریبا هر آنچه نیاز دارید در اختیار شما قرار داده شده است. بنابراین پیچیدگی پیاده‌سازی مطرح نیست. مهم است که هر مشاهده را با منطق و برداشت خود از نحوه عملکرد شبکه‌های عصبی و توسعه LSTM تحلیل کنید. تا جای ممکن، مشاهدات را با دلایلی شهودی نیز توجیه کنید.

5- قلم نویسنده

از کتاب‌های موجود در دادگان، سه کتاب دلخواه انتخاب کنید. یک مدل زبانی مبتنی بر LSTM با جزئیات دلخواه خود بسازید و هر بار آن را بر روی یکی از این چهار کتاب آموزش دهید (نیازی به ارزیابی نیست). حال به کمک هر یک از این چهار مدل آموزش دیده، یک جمله به طول دلخواه (یا متغیر) تولید کنید. برای تولید جملات منطقی‌تر، می‌توانید راهکارهای متفاوتی پیش بگیرید. به عنوان مثال ممکن است تصمیم بگیرید برای انتخاب لغت جدید، از argmax استفاده نکنید یا موارد خاصی از کلمات را ممنوع کنید (این راهکار پیشنهاد نمی‌شود). بهتر است در این تولید، اجازه‌ی تولید کلمه‌ی <UNK> (یا مواردی از این جنس که در پیش‌پردازش اضافه کردید) ندهید. اگر ترجیح می‌دهید، بخشی از ابتدای جمله را به مدل ورودی بدهید (برای هر سه کتاب ورودی یکسانی بدهید). ممکن است در جملات تولید شده، فقط شاهد stop words و کلمات پرکاربرد، بدون توجه به جایگاه آن‌ها باشید که خبر بدی خواهد بود. در صورت روبرو شدن با این شرایط، این مشاهده را توجیه کرده و با تغییر در مدل، (تا حدی که ممکن است) مانع این موضوع شوید.

عملکردی مشابه آنچه خواسته شده، از مدل ساده‌ای که برای طرح تمرین پیاده‌سازی شده را می‌بینیم:

+ کتاب:

crocodiles-spirit

+ جمله تولید شده:

she had lost part of the northern territory.

+ متنی که عبارت بالا از آن الهام گرفته شده است:

Author's Note

This is a novel set in Australia's Northern Territory, a place where I lived and worked for four decades; including in small towns, aboriginal communities, cattle stations and among remote, rugged and beautiful natural places for which it is famous, places with names like Uluru and Kakadu.

عبارت بالا، تنها بخشی از کتاب است که به Northern Territory اشاره دارد. نحوه ادراک مدل جالب نیست؟

6- ملاحظات

- تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_HW3_StudentID تحویل داده شود.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. تمرین‌هایی که به صورت عکس در سایت بارگذاری شوند، ترتیب اثر داده نخواهند شد. گزارش نهایی خود را حتماً به صورت PDF در سایت درس بارگذاری نمائید.
- گزارش به صورت تایپ شده در قالب PDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش و نتایج و تحلیل‌ها باشد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تأخیر تحویل تمرین تا یک هفته، هر روز ۱۰ درصد است.
- توجه کنید این تمرین باید به صورت تک‌نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد. در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر داده شده و موضوع به استاد گزارش می‌گردد.
- در صورت بروز هرگونه مشکل، با این ایمیل[16] در ارتباط باشید.
- مهلت تحویل بدون جریمه: ۱۲ اردیبهشت ۱۴۰۰
- مهلت تحویل با تأخیر، با جریمه: ۱۹ اردیبهشت ۱۴۰۰
- اگر مشتاق دریافت بازخورد متنی حاصل از بررسی گزارش خود هستید، این موضوع را به همراه آدرس ایمیل خود در انتهای گزارش بیاورید.

[16] behzad.shayegh@ut.ac.ir