به نام خدا

دانشکده مهندسی برق و کامپیوتر

درس یادگیری عمیق

تمرین سری سوم

در این تمرین هدف پیاده‌سازی تسک تحلیل احساسات (Sentiment Analysis) است و در آن از شبکه‌های بازگشتی (RNN) و شبکه‌ی BERT استفاده می‌کنیم. مجموعه داده‌ی مورد استفاده در این تمرین، دیتاست Sentiment140 خواهد بود که از لینک زیر می‌توانید آن را دریافت کنید و با جزئیات آن آشنا شوید:

http://help.sentiment140.com/for-students

کتابخانه‌ی مورد استفاده در این تمرین، PyTorch خواهد بود.

پیش پردازش

ابتدا لازم است تا یک پیش پردازش روی دیتاست انجام دهید و داده‌ها را آماده‌ی استفاده کنید. بدین منظور، کارهای زیر را قبل از وارد شدن به دو بخش این تمرین انجام دهید و داده‌های نهایی را ذخیره کنید:

1. هر یک از داده‌های این دیتاست، یک توییت است. لازم است تا هشتگ‌ها و منشن‌ها را با یک عبارت یا کلمه جایگزین کنید.
2. توییت‌های حاوی لینک را به طور کلی حذف کنید. استفاده از لینک‌ها در مدل، ممکن است باعث یادگیری پترن‌های نادرست شود و دقت مدل کاهش یابد.
3. تمامی علائم نگارشی را از توییت‌ها حذف کنید.
4. برچسب داده‌ها را به ۰ (منفی)، ۱ (خنثی) و ۲ (مثبت) تغییر دهید.

بخش اول

در این بخش، به کمک شبکه‌های بازگشتی از نوع LSTM، مدل را طراحی می‌کنیم و آموزش می‌دهیم.

در ابتدا، قصد داریم تا به هر کلمه به کمک یک word embedding (در حالت ساده، یک عدد یکتا یا یک بردار one-hot) نسبت دهیم. در این بخش، از glove42b با بعد ۳۰۰ استفاده می‌کنیم. این embedding، در واقع هر لغت را به یک بردار ویژگی با بعد ۳۰۰ مپ می‌کند. از لینک زیر می‌توانید آن را دریافت کنید:

http://nlp.stanford.edu/data/glove.42B.300d.zip

نکته‌ی دیگر، طول هر یک از ورودی‌های مدل (تعداد کلمات هر ورودی) است. لازم است تا با استفاده از padding مناسب، طول ورودی را یکسان و برابر با ۲۸۰ کنید.

1. ساختار مدل را به صورت زیر در نظر بگیرید:
   در ابتدا ورودی به یک شبکه‌ی LSTM یک‌طرفه با یک لایه و بعد مخفی ۱۵۰ داده می‌شود و پس از آن، خروجی این لایه به یک لایه‌ی خطی با ۳ خروجی داده می شود و در نهایت، تابع Softmax روی خروجی لایه‌ی خطی اعمال می‌شود و خروجی نهایی مدل ساخته می‌شود.
   (در واقع، لایه‌ی LSTM گویا یک لایه‌ی استخراج ویژگی از ورودی است)

همانطور که بالاتر گفته‌شده‌است، ابتدا ورودی خام را با استفاده از glove42b به بردارهای ویژگی تبدیل کنید (ممکن است یک لغت در این word embedding وجود نداشته باشد. در این حالت، به جای بردار آن لغت، میانگین تمامی بردارهای ویژگی موجود در glove42b را در نظر بگیرید) و سپس طول جملات (تعداد بردارهای ویژگی هر ورودی) را یکسان‌سازی کنید (می‌توانید این کارها را در بخش پیش‌پردازش نیز انجام دهید و این داده‌های نهایی را که بخش اول از آن‌ها استفاده می‌کنیم، ذخیره کنید). حال، ورودی ما آماده‌ی استفاده در مدل است.

مدل را آموزش دهید و نمودار خطای داده‌های آموزش و تست و دقت داده‌های تست مدل در طول زمان و همچنین ماتریس درهم‌ریختگی (Confusion Matrix) در انتهای آموزش را گزارش کنید.

پارامترهای شبکه را نیز در این تمرین به صورت زیر در نظر بگیرید (مقدار پارامترهایی مانند Batch Size و Learning Rate را دلخواه در نظر بگیرید)

Optimizer	Adam
Loss Function	Cross entropy

2. در مدل بخش قبل، از LSTM دو طرفه استفاده کنید و تغییرات مورد نیاز در باقی بخش‌های شبکه را نیز اعمال کنید (دقت کنید که خروجی یک بردار با طول ۳ خواهد بود و تغییر نخواهد کرد) و تفاوت آن با بخش قبل را توضیح دهید و نتایج خواسته‌شده در بخش قبل را پس از آموزش مدل جدید گزارش کنید.

3. LSTM یک‌طرفه‌ی بخش اول را با یک ساختار هرمی (Pyramid) با ۴ سطح جایگزین کنید و تغییرات مورد نیاز در باقی بخش‌های شبکه را نیز اعمال کنید. ساختار هرمی در این سوال

بدین صورت است که در هر سطح، ورودی هر سلول از اتصال (concatenate) خروجی دو سلول سطح پایین‌تر ساخته می‌شود. بعد مخفی پایین‌ترین سطح را ۶۴ در نظر بگیرید و نتایج را گزارش کنید (در صورت طولانی شدن زمان آموزش به میزان قابل توجهی، از نصف داده‌های آموزش برای آموزش مدل استفاده کنید)

برای آشنایی بیشتر با ساختار هرمی، می‌توانید به بخش مربوط به Listener از مقاله‌ی زیر مراجعه کنید.

https://arxiv.org/pdf/1508.01211

بخش دوم

در این بخش قصد داریم تا به کمک یک مدل BERT از قبل آموزش دیده (pre-trained)، تحلیل احساسات انجام دهیم و آن را روی داده‌های خود fine-tune کنیم و نتایج را با بخش قبل مقایسه کنیم.

برای آشنایی با ساختار BERT می‌توانید از لینک‌های زیر استفاده کنید.

http://jalammar.github.io/illustrated-bert/

https://arxiv.org/abs/1810.04805

https://towardsdatascience.com/bert-classifier-just-another-pytorch-model-881b3cf05784

در این بخش از کتابخانه‌ی pytorch_pretrained_bert استفاده می‌کنیم.

https://pypi.org/project/pytorch-pretrained-bert/

1. ساختار مدل را به صورت زیر در نظر بگیرید.

در ابتدا ورودی (پس از استفاده از tokenizer که در ادامه توضیح داده خواهد شد) به مدل BERT داده می‌شود و سپس خروجی آن به یک لایه‌ی خطی با ۳ خروجی داده می‌شود و در نهایت تابع Softmax روی آن اعمال می‌شود.

مدل گفته‌شده را تا حداکثر ۳ epoch روی داده‌ها آموزش دهید و نتایج را گزارش کنید و با نتایج به دست آمده از بخش اول، مقایسه کنید.

نکات مربوط به بخش دوم:

- وظیفه‌ی tokenizer مشابه یک word embedding است. در واقع، tokenizer هر جمله را به توکن‌های از پیش تعریف شده تقسیم می‌کند و سپس به هر کدام از آن‌ها، یک آیدی نسبت می‌دهد. این آیدی‌ها در نهایت، ورودی مدل BERT خواهند بود.
- همانند بخش قبل، لازم است که طول ورودی‌ها را یکسان کنید و این طول یکسان را ۳۰۰ در نظر بگیرید (بدین منظور، پس از استخراج آیدی هر توکن موجود در یک توییت، به انتهای آن به تعداد لازم صفر اضافه کنید و یا در صورت بزرگ‌تر بودن طول ورودی، تنها از ۳۰۰ توکن اول آن استفاده کنید).
- در این بخش از مدل و tokenizer آموزش داده شده‌ی bert-base-uncase استفاده کنید. این مدل دارای ۱۲ لایه با بعد مخفی ۷۶۸ است.
- پارامترهای آموزش شبکه را مشابه بخش اول در نظر بگیرید.

نکات کلی تمرین:

- در صورت مشاهده هر گونه مشابهت کد بین هر دو دانشجو، نمره تمرین هر دو دانشجو صفر لحاظ خواهد شد.
- در صورت مشاهده هر گونه مشابهت کد با کد های موجود در صفحات اینترنتی، نمره تمرین صفر لحاظ خواهد شد. اگر بخشی از کد را از کد آماده اینترنتی استفاده می کنید که جزو قسمت های اصلی تمرین نمی باشد، حتما باید لینک آن در گزارش و کد ارجاع داده شود.
- توجه نمایید که نیمی از نمره تمرین مربوط به گزارش می‌باشد. لازم به ذکر است رعایت اصول نگارشی حائز اهمیت است.
- در نوشتن گزارش، لحاظ جزییات نوشتن گزارش الزامی است. مانند موارد زیر:
  - ارجاع دادن به مطالب و اشکالی که از مقاله و وبسایت ها گرفته شده است.
  - توضیح اشکال و جداول در caption
  - نوشتن فرمول و قرار ندادن عکس مربوط به فرمول
  - ارجاع به شکل و جدول در متن گزارش
  - نوشتن نتایج شبیه سازی ها به صورت جدولی و شکل (از قرار دادن عکس نتیجه اجرای کد پرهیز شود)
  - درست بودن متن از نظر قواعد دستور زبانی و نگارشی
  - موارد تکمیلی در فایل template توضیح داده شده اند.
- گزارش تمرین را حتما به صورت PDF و در کنار کدهای تمرین در سایت درس آپلود نمائید.
- نحوه نامگذاری به صورت studentnumber_homeworknumber.pdf می باشد.
- برای پیاده سازی می توانید از محیط colab استفاده نمایید.
- هرگونه پرسش پیرامون تمرین را با ایمیل های aliparchekan@gmail.com و amir.karimi6610@gmail.com مطرح نمایید.

موفق باشید