به نام پروردگار

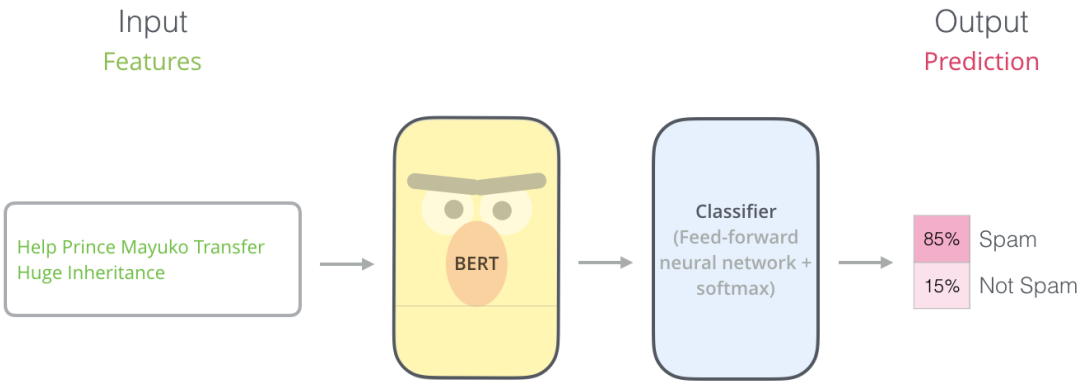
پردازش زبان طبیعی تمرین اول

موعد تحویل :

میلاد صیادآموز (sayad.mehrdad@gmail.com)

در این تمرین هدف داریم با استفاده از روش های بردارهای تعبیه شده متنی[1] ، کارهای پردازش زبان مثل تحلیل احساس و تشخیص اسپم را انجام دهیم. دو مدل از پیش آموزش دیده شده که بسیار معروف هستند عبارتند از مدل ELMo و مدل BERT.

هدف اصلی این دو مدل از پیش آموزش داده شده این است که بردارهای معنایی مرتبط با پردازش و داده های زمینه کاری تولید کند. ما قصد داریم این دو مدل را طوری آموزش دهیم و در شبکه مورد نظر خود قرار دهیم که مساله ما را حل کند. نمونه ای از این شبکه به صورت زیر است.



شکل ۱ استفاده از مدل BERT درتشخیص اسپم[2]

شکل ۱ رده بندی را نشان می دهد که با استفاده از مدل BERT برای تشخیص پیام های اسپم آموزش دیده شده، وزن های مدل شبکه عصبی BERT براساس داده های آموزش اسپم یادگیری شده است لذا برای این کار مناسب است.

---

[1] contextual embedding

[2] http://jalammar.github.io/images/BERT-classification-spam.png

هدف ما این است طبق شکل ۱ شبکه ای تعبیه کنیم که با استفاده از دو مدل از پیش آموزش داده شده متنی، پاسخگوی تشخیص اسپم و تحلیل احساس باشد.

در این تمرین دو دیتاست در اختیار شما قرار گرفته است که برای آموزش شبکه خود و ارزیابی از آن دو می توانید استفاده کنید.

سوال 1 تشخیص اسپم

با استفاده از دیتاست sms-spam-collection-dataset.zip[3] که حاوی متن پیام ها است که تعدادی از این پیام ها اسپم هستند. در این قسمت می خواهیم همانند شکل ۱ با استفاده از دو مدل از پیش آموزش داده شده و یک شبکه Feed Forward یک لایه ای استفاده کنیم و بعد از آموزش مدل بدست آمده را ارزیابی کنیم. برای پاسخگویی به این سوال نیازمند است که مراحل زیر را طی کنیم.

لینک دانلود دیتاست:

https://drive.google.com/open?id=1cNExwm8MLE1a2ZKtRzTOfl8Q7pQS4Lds

1) دیتاست را به دوبخش ۲۰ درصدی تست و ۸۰ درصدی آموزش به صورت تصادفی تقسیم کنید.
2) پیش پردازش های مناسب (tokenization, remove stop words, remove punctuations)
3) اضافه کردن مدل از پیش آموزش شده به عنوان یکی از لایه ها شبکه
4) اضافه کردن لایه شبکه feed forward برای طبقه بندی
5) اضافه کردن لایه softmax و طراحی Decoder برای خروجی

برای هر دو مدل BERT[4] و Elmo[5] این شبکه را طراحی کنید پارامترهای شبکه را به صورت زیر تنظیم کنید.

---

[3] SMS Spam Collection Dataset | Kaggle
[4] https://tfhub.dev/tensorflow/bert_en_cased_L-24_H-1024_A-16/1
[5] https://tfhub.dev/google/elmo/3

(بیشترین تعداد ورودی = ۱۲۸ کلمه ،تعداد نورون های لایه شبکه feed forward = اندازه بردار خروجی مدل Bert یا Elmo، نرخ یادگیری = ۰.۰۰۰۲ و (batch size) اندازه دسته ها = ۳۲)

الف)برای ۱۰ ، ۲۰ و ۵۰ تکرار (epoch) شبکه را آموزش دهید و پس از آموزش نمودار تغییرات loss گزارش کنید و برای قسمت تست precision،recall، F1، AUC و Accuracy را گزارش کنید .آیا تعداد تکرار در دقت رده بند موثر است؟

ب)تاثیر پیش پردازش را روی آموزش و دقت رده بند طراحی شده بررسی کنید.

ج) برای طبقه بند تشخیص اسپم کدام معیار مهم تر است؟ precision یا recall؟ علت را توضیح دهید.

سوال ۲ تحلیل احساس

با استفاده از دیتاست aclImdb[6] که نظرات کاربران در مورد فیلم های سینمایی است که توسط گروه پردازش زبان استنفورد منتشر شده است، مراحل سوال ۱ را برای این قسمت هم پیاده و طراحی کنید.(فقط در این سوال نیازی به جداسازی داده آموزش و تست در مرحله ۱ نمی باشد زیرا داده های آموزش و تست جدا می باشند)

الف)برای ۱۰ ، ۲۰ و ۵۰ تکرار (epoch) شبکه را آموزش دهید و پس از آموزش نمودار تغییرات loss گزارش کنید و برای قسمت تست precision،recall، F1، AUC و Accuracy را گزارش کنید. آیا تعداد تکرار در دقت رده بند موثر است؟

ب)تاثیر پیش پردازش را روی آموزش و دقت رده بند طراحی شده بررسی کنید.

سوال ۳ (امتیازی)

به یکی از سوالات زیر پاسخ دهید.

۱) در شبکه طراحی شده در قسمت های پیشین از شبکه Feed Forward استفاده کردیم اگر شبکه پیشنهاد دارید آن را پیاده کنید در غیر این صورت شبکه fully connected را پیاده کنید و نتایج به دست آمده را گزارش کنید و با قسمت های قبلی مقایسه کنید.

---

[6] http://ai.stanford.edu/~amaas/data/sentiment/

۲) آیا تعداد لایه های شبکه Feed Forward در دقت رده بند موثر است؟ برای اثبات این موضوع می‌توانید شبکه با ۲، ۵ و ۷ لایه را آموزش دهید و تغییرات در معیارها را گزارش کنید و با مقایسه آن ها به سوال جواب دهید.

لطفا به قواعد حل تمرین که در CECM قرار داده شده است توجه کنید.

* استفاده از هر کتابخانه‌ای مجاز است.

* پیشنهاد است از google colab استفاده کنید.(آدرس دانلود مدل های BERT و Elmo در پانوشت قرار دارد که قابل استفاده در google colab می باشند)