به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

Natural Language Processing

تمرین 4

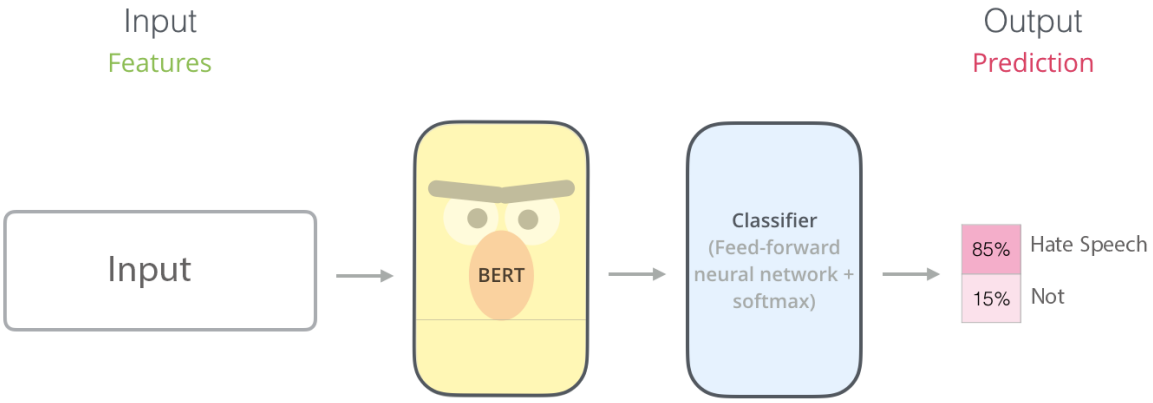
اردیبهشت ماه 1400

فهرست

مقدمه

در این تمرین قصد داریم با استفاده از روش های بردارهای تعبیه شده متنی، کارهای پردازش زبان مثل شناسایی تنفر را بر روی داده های تمرین دوم به دست آوریم. دو مدل از پیش آموزش دیده شده که مبتنی بر Transfomer ها هستند عبارتند از مدل XLNet و BERT.

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



شکل ۱ استفاده از مدل **BERT** درتشخیص تنفر

مجموعه دادگان در نظر گرفته شده از شبکه اجتماعی توئیتر جمع‌آوری گردیده و کلاس‌های دادگان به صورت دوتایی می‌باشد (هر پیام، تنفر می‌باشد یا نه) که در تمرین های گذشته , با آن آشنا شده اید.

دادگان	داده‌های آموزشی		داده‌های تست	
	تنفر	غیر تنفر	تنفر	غیر تنفر
Hate Eval	10000		3000	
	4210	5790	1260	1740

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 تنفر و تحلیل احساس باشد

سوال 1

با استفاده از دیتاست موجود که حاوی متن پیام ها است , میخواهیم همانند شکل ۱ با استفاده از Transformer ها, دو مدل از پیش آموزش داده شده و یک شبکه یک لایه ایی Feed Forward از کتابخانه pytorch 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.

1. پیش پردازش های مناسب را انجام دهید.
2. اضافه کردن مدل از پیش آموزش شده به عنوان یکی از لایه ها شبکه
3. اضافه کردن لایه شبکه feed forward برای طبقه بندی
4. اضافه کردن لایه softmax برای ایجاد خروجی

برای هر دو مدل BERT و XLNet این شبکه را طراحی کنید . پارامترهای شبکه را نیز می توانید در صورت لزوم به صورت زیر تنظیم کنید.

بیشترین تعداد ورودی ۱۲۸ کلمه ، تعداد نورون های لایه شبکه feed forward = اندازه بردار خروجی مدل Bert یا XLNet ، نرخ یادگیری = 0.0002 , size batch یا اندازه دسته ها = ۳۲

سوال 2

با تعداد تکرار مختلف (Epochs) شبکه را آموزش دهید و پس از آموزش نمودار تغییرات loss را گزارش کنید و برای قسمت تست precision ،recall، F1 و Accuracy را گزارش کنید .آیا تعداد تکرار در دقت رده بند موثر است؟ نتایج بدست آمده را تحلیل کنید.

سوال 3

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

***میزان تعداد تکرار و یا بدست آوردن دقت یا Accuracy خاصی در این تمرین ملاک نیست , بلکه روند یادگیری صحیح مدل و تحلیل های مبتنی بر آنها اهمیت دارد.**

ملاحظات (حتماً مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_CA4_StudendID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد. گزارش نهایی خود را حتماً به صورت PDF در سایت درس بارگذاری نمائید.
- گزارش به صورت تایپ شده در قالبPDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. سیاست تحویل تمرین و پروژه ها اینچنین است که هر روز تاخیر 10٪ کسر نمره دارد و حداکثر 7 روز تاخیر پذیرفته می‌شود (حداکثر 70٪ کسر نمره دارد)
- 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 و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با دستیاران آموزشی زیر در ارتباط باشید:

Beheshti.7676@gmail.com

Yaser.abbaszadeh@ut.ac.ir

مهلت تحویل بدون جریمه: یکشنبه 26 اردیبهشت 1400

مهلت تحویل با تأخیر: یکشنبه 2 خرداد 1400

موفق باشید.