



هوش مصنوعی

تمرین شماره 4

مدرسین: دکتر فدایی و
دکتر یعقوبزاده

طراحان: مهرداد لیویان، محمد شهاب شرافت، جواد
کاویان

مهلت تحویل: 1404/10/6

مقدمه

در این تمرین، به پیاده سازی یک شبکه عصبی پیچشی¹ بر روی داده متنی می پردازید. هدف اصلی این تمرین، کسب تجربه عملی در پیاده سازی و آموزش یک شبکه عصبی است. توجه کنید که در این تمرین، می بایست از فریم ورک PyTorch استفاده کنید و با فرمت استاندارد برنامه نویسی در آن، آشنا شوید. برای این امر، به نوتبوک آموزشی که ضمیمه تمرین است مراجعه کنید. در صورت نیاز می توانید از منابع آموزشی دیگر نیز استفاده کنید.

PyTorch Warmup

در این نوتبوک خودمان دیتاست ساده و معروف XOR را می سازیم که از منطق xor به همراه مقداری نویز پیروی می کند. این دیتاست به دلیل این که با یک خط ساده قابلیت جداسازی ندارد، از اولین دیتاست هایی است که در دنیای یادگیری ماشین لرنینگ به آن پرداخته شده است. در این بخش شما وظیفه دارید با مدل شبکه عصبی سبکی که طراحی کردیم، در این دیتاست classification را انجام دهید. بدین منظور لازم است کد بخش حلقه Training را کامل کنید و در نهایت نتیجه را بررسی کنید.

شرح مسئله اصلی

مسئله ای که در این تمرین با آن روبرو هستیم، مسئله تحلیل احساسات² است که یکی از مسائل معروف در پردازش زبان طبیعی است. دیتاستی که در این تمرین با آن سروکار داریم، دیتاستی است که اسنپ فود³ آن را منتشر کرده و از نظرات مشتریان در مورد غذاهای خریداری شده تشکیل شده است. هر سطر دیتاست، شامل دو ستون است که یک ستون آن، رشته متن نظر مشتری است و ستون دیگر آن، حالت آن جمله است که به صورت HAPPY یا SAD می باشد. در این تمرین، از شما انتظار می رود یک شبکه عصبی پیچشی آموزش دهید و با استفاده از آن، حالت جملات را پیش بینی کنید.

¹ Convolutional Neural Network

² Sentiment Analysis

³ SnappFood

ابتدا می بایست دیتاست ها را لود کنید. توصیه می شود برای تمیزی کد، یک کلاس دیتاست بنویسید که از کلاس Dataset کتابخانه PyTorch ارث بری می کند.

مرحله بعدی، پیش پردازش متن خام است. از شما انتظار می رود موارد زیر را انجام دهید:

1. جایگزین کردن کاراکتر های <ي> و <ك> با <ی> و <ک>
2. جایگزین کردن اعداد انگلیسی با فارسی
3. حذف لینک ها و mention ها از متن
4. حذف white space های اضافی
5. حذف stop word ها

می توانید جهت اطمینان هر نویسه ای که جزو حروف، اعداد یا علائم رایج نگارشی فارسی نباشد را حذف کنید. همچنین برای دسترسی به stop word های زبان فارسی، می توانید از این [لینک](#) استفاده کنید.

مرحله بعدی، embedding است. توجه کنید که می بایست داده متنی را تبدیل به عدد کنیم تا مدل ما آن را بفهمد. برای این کار، می توان از مدل های از پیش آموزش دیده استفاده کرد که می توانند کلمات را تبدیل به وکتور کنند. مدل مورد استفاده در این تمرین، مدل fasttext است. ابتدا مدل را لود کنید. سپس با استفاده از آن، embedding کلمات جمله را بدست آورید و با عملیات concatenation، یک embedding برای جمله بسازید. مشکلی که اینجا وجود دارد، این است که طول جملات متفاوت است. برای این کار می توان از عملیات های padding و truncate استفاده کرد؛ در مورد این عملیات ها در گزارش کار خود توضیح دهید.

حال می بایست یک شبکه CNN تعریف کنید و آن را روی دیتاست مطرح شده آموزش دهید. یک معماری پیشنهادی، معماری زیر است :

- Conv1D(input_dimension, 128, kernel_size=3, padding=1)
- MaxPool1d(kernel_size = 2)
- Linear(input_dimension_linear, 2)

برای جلوگیری از overfitting در شبکه های عصبی، می توان از تکنیک drop out استفاده کرد. در این مورد تحقیق کرده و در معماری خود استفاده کنید.

در نهایت پس از تعریف مدل، مدل را آموزش دهید. از CrossEntropyLoss به عنوان تابع هزینه استفاده کنید. مدل را حداقل به مدت 10 اپیاک آموزش دهید.

پس از آموزش مدل، دقت آن را روی داده تست ارزیابی کرده و گزارش کنید.

نکات پایانی

- دقت کنید که کد شما باید به نحوی زده شده باشد که نتایج قابلیت بازتولید داشته باشند.
- توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. حجم توضیحات گزارش شما هیچ گونه تاثیری در نمره نخواهد داشت و تحلیل شما بیشترین ارزش را دارد.
- سعی کنید از پاسخ‌های روشن در گزارش خود استفاده کنید و اگر پیش‌فرضی در حل سوال در ذهن خود دارید، حتما در گزارش خود آن را ذکر نمایید.
- فایل‌های خود را در قالب یک فایل فشرده با فرمت `AI_CA3_[stdNum].zip` در سامانه ایلرن بارگذاری کنید. به طور مثال `AI_CA3_810102123.zip`.
- محتویات پوشه باید شامل گزارش و کدهای شما باشد.

موفق باشید.