شبکه‌های عصبی بازگشتی

سوال ۱

در این تمرین، هدف بررسی کاربرد مدل‌های بازگشتی و ترکیبی برای طبقه‌بندی ایمیل‌های اسپم به زبان فارسی است. در این پروژه، شما با استفاده از یک مجموعه‌داده از ایمیل‌های فارسی، عملکرد مدل‌های مختلف را پیاده‌سازی، مقایسه و تحلیل خواهید کرد.

مجموعه‌داده:
از مجموعه‌داده زیر برای طبقه‌بندی ایمیل‌ها به دو دسته «اسپم» و «عادی» استفاده نمایید:

https://www.kaggle.com/datasets/mohamad1dehqani/persian-spam-email

این مجموعه شامل ایمیل‌های فارسی با برچسب «اسپم» و «نرمال» است. در ابتدا داده‌ها را به شیوه مناسب پیش‌پردازش کرده (از جمله پاک‌سازی متن، توکن‌سازی، حذف علائم نگارشی و...) و سپس آن‌ها را به دنباله‌های عددی تبدیل نمایید.

۱ - تمام مراحل پیش‌پردازش را بطور کامل انجام داده و تمام مراحل را بهمراه دلیل انتخاب آن‌ها در فایل گزارش ذکر نمایید.

۲ - مدلی ساده شامل لایه تعبیه [۱]، یک لایه بازگشتی و یک لایه کاملا متصل[۲] طراحی و پیاده‌سازی کنید. مدل را آموزش دهید و معیارهای Accuracy، Precision، Recall و F1-score را گزارش نمایید. همچنین منحنی‌های خطا را برای داده‌های آموزش و اعتبارسنجی رسم و تحلیل نمایید.

۳ - همین ساختار را با جایگزینی لایه بازگشتی با یک لایه حافظه کوتاه‌مدت بلند[۳] پیاده‌سازی کرده و عملکرد آن را با حالت قبل مقایسه و تحلیل نمایید.

۴ - در این حالت، حساسیت مدل به پارامترهای زیر را آزمایش و تحلیل نمایید. توجه نمایید در هر مورد حداقل ۳ مقدار مختلف از پارامتر مورد نظر باید آزمایش شود.

[۱] Embedding
[۲] Fully Connected
[۳] LSTM

سوال ۱ ادامه در صفحه بعد...

- طول دنباله ورودی
- اندازه بردار تعبیه
- ابعاد بردار حالت[۴] شبکه حافظه کوتاه‌مدت بلند
- استفاده/عدم استفاده از Dropout

نتایج را به صورت جدول و نمودار ارائه کرده و تحلیل نمایید که کدام پارامترها حساسیت بیشتری دارند.

۵ - تأثیر استفاده از مدل حافظه کوتاه‌مدت بلند دوطرفه [۵] را بررسی کرده و نتایج را با حالت یک‌طرفه مقایسه نمایید.

۶ - مدلی ترکیبی طراحی کنید که پس از لایه تعبیه، از یک یا چند لایه پیچشی یک بعدی [۶] استفاده کرده و خروجی آن را به یک لایه حافظه کوتاه‌مدت بلند متصل نمایید. عملکرد این مدل را با مدل‌های قبلی مقایسه و تحلیل کنید. در این بخش تاثیر افزایش یا کاهش تعداد لایه‌های پیچشی بر عملکرد مدل را نیز مقایسه و تحلیل نمایید. توجه نمایید هر لایه پیچشی با یک لایه زیرنمونه‌برداری [۷] ترکیب میگردد.

۷ - با استفاده از داده‌های آزمون، نمونه‌هایی از ایمیل‌هایی که مدل به درستی و به اشتباه طبقه‌بندی کرده است انتخاب نموده و تحلیل کنید که چرا مدل دچار خطا شده است. آیا علت در ویژگی‌های متن بوده یا محدودیت مدل؟

در انجام تمرینات به نکات زیر توجه فرمایید.

۱ - پیاده‌سازی‌های کامپیوتری را به زبان برنامه‌نویسی پایتون و با بهره‌گیری از چارچوب کاری PyTorch انجام دهید.

۲ - بخش عمده‌ای از نمره تمرینات به گزارش تمرین اختصاص دارد و ارسال برنامه‌ها بدون گزارش فاقد ارزش است. در تهیه گزارش دقت نمایید که تمام اطلاعات، تصاویر و نمودارهای مورد نیاز برای اثبات پاسخ‌ها مبتنی بر آزمایشات خواسته شده در تمرین، بطور کامل و دقیق ذکر شده باشند.

[۴] Hidden state layer
[۵] Bidirectional LSTM
[۶] 1D convolutional layer
[۷] Down sampling layer

۳ - مطابق قوانین دانشگاه هر نوع کپی‌برداری و اشتراک کار دانشجویان غیرمجاز است. در صورت مشاهده چنین مواردی، با طرفین شدیدا برخورد خواهد شد.

۴ - استفاده از کدها و توضیحات اینترنت یا کدها و توضیحات تولید شده با مدل‌های هوش مصنوعی به منظور یادگیری بلامانع است، اما کپی‌برداری و انجام تمرینات توسط این ابزارها غیرمجاز است. در صورتی که از چنین ابزارهایی بهره می‌گیرید، حتما به تمام جزئیات و نکات مرتبط با پاسخ‌ها مسلط باشید، در غیر اینصورت نمره کل تمرین را از دست خواهید داد.

۵ - مجموعه‌داده‌های مورد استفاده را به جز در مواردی که صریحا در صورت سوال ذکر شده باشد، حتما قبل از استفاده بطور تصادفی به سه بخش آموزشی (۷۰ درصد)، آزمون (۲۰ درصد) و اعتبارسنجی (۱۰ درصد) تقسیم نمایید.

۶ - در صورت نیاز می‌توانید سوالات خود را در خصوص پروژه از تدریسیارهای درس، به ایمیل زیر ارسال نموده یا در گروه «بله» مطرح نمایید.

ann.ceit.aut@gmail.com

۷ - فایل‌های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_HW05.zip تا تاریخ ۱۴۰۴/۰۳/۰۲ فقط در بخش مربوطه در سایت درس بارگذاری نمایید. توجه نمایید، هر روز تاخیر منجر به کسر ۱۰ درصد از نمره پروژه می‌شود.