تصاویر پاتولوژی

تعریف مسئله

در این سوال قصد داریم با دیتاست patch_camelyon کار کنیم. دیتاست patch_camelyon شامل دو کلاس از تصاویر هیستوپاتولوژی میباشد. ابعاد تصاویر در این دیتاست `96x96` است که در دو کلاس `سالم(0)` و `سرطانی(1)` تقسیم بندی شده اند . شما باید به کمک دیتای train که در اختیار دارید یک مدل KNN classifier آموزش دهید و دقت مدل را بر روی داده تست بدست آورید. برای این کار باید ابتدا یک مدل autoencoder را به کمک دیتا train آموزش دهید. سپس به کمک لایه latent این مدل ، امبدینگ تصاویر موجود در دیتاست را بدست آورید. در انتها به کمک امبدینگ های بدست آمده، یک مدل KNN classifier را آموزش دهید و دقت مدل را بر روی دیتای test گزارش دهید.

پاکسازی داده

یکی از چالش هایی که با آن روبرو هستید، تمیز نبودن دیتای train است. در واقع دیتای train فقط شامل تصاویر هیستوپاتولوژی نیست و تعدادی تصویر از تصاویر دیتاست imageNet نیز در این مجموعه قرار دارند. شما باید در قدم اول به کمک روشی این تصاویر را از مجموعه train حذف کنید. روش و ایده شما برای حذف این تصاویر باید اصولی و درست باشد و حذف تصاویر به صورت دستی و با نظارت انسانی قابل قبول نیست.

دیتاست

دیتاست این مسئله را میتوانید از این لینک دریافت کنید. دیتا بخش train دیتای مدنظر برای پاکسازی و آموزش مدل Autoencoder است.

خروجی

برای این مسئله باید دو فایل با نام های clean_data.txt و output_model.txt را در قالب یک فایل zip آپلود کنید . فایل clean_data.txt شامل نام تصاویری است که به عنوان داده پرت در دیتاست شناسایی کردید. در هر سطر از این فایل تنها نام فایل تصاویر ثبت میشود. مانند نمونه زیر:

خروجی نمونه

```
12412.png
03423245.png
17756.png
```

فایل output_model.txt شامل کلاس هر یک از دادگان تست است.در واقع هر سطر این فایل فقط حاوی عدد 0 یا 1 است که نشان دهنده کلاس تصویر تست است. ترتیبی که قرار است بر اساس آن کلاس تصاویر را مشخص کنید بر اساس این فایل باشد. نمونه خروجی مدنظر برای این بخش:

خروجی نمونه

```
0
1
1
1
0
1
0
0
.
.
.
```

نکات

۱. استفاده از کتابخانه هایی مثل sklearn و همچنین مدل های pretrain مجاز است.

۲. در بخش پیاده سازی Autoencoder مجاز به استفاده از Autoencoder های کانولوشنی و غیر کانولوشنی هستید.