یادگیری ماشین

تمرین سری ششم

عمادالدین رستمیان
403206479

پاسخ سوال ۱

در یادگیری خودنظارتی، وظیفه پیش‌متن یک وظیفه کمکی است تا به مدل کمک کند که ویژگی های موثری را از داده های بدون برچسب یاد بگیرد. در واقع هدف این است که مدل بتواند با استفاده از برچسب‌های تولید شده به صورت خودکار (بدون نیاز به دخالت انسان) این وظیفه را انجام دهد و در نتیجه آن، مدل representation هایی را یاد بگیرد که در انجام وظایف بعدی نیز عملکرد خوبی داشته باشند.
وظایف پیش‌متن معمولاً ساده هستند، اما مدل را مجبور می‌کنند که الگوها یا ساختارهای معنی‌داری را از داده‌های ورودی استخراج کند.

الف) پیش‌بینی چرخش: در این وظیفه پیش‌متن تصاویر به صورت رندوم و با زوایای مشخص (مانند ۹۰، ۱۸۰ یا ۲۷۰ درجه) چرخیده می‌شوند و مدل باید زاویه چرخش را برای هر تصویر پیش‌بینی کند. آموزش مدل برای این وظیفه باعث میشود تا مدل جهت گیری های فضایی و ساختار کلی اجسام درون تصاویر را یاد بگیرد.

ب) رنگ‌آمیزی: با تغییر رنگ پیکسل های تصویر به رنگ های دیگر یا به grayscale مدل برای پیش‌بینی درست رنگ پیکسل ها یا نواحی مختلف تصویر آموزش می‌دهیم. این وظیفه پیش‌متن مدل را مجبور به یادگیری توزیع رنگ های اجسام، روابط معنایی بین آن ها و بافت های مختلف تصویر می کند.

ج) حل پازل: در این تسک تصویر به patch های مختلف تقسیم می‌شود. این پچ ها به صورت رندوم shuffle می‌شوند و مدل برای بازسازی تصویر اصلی آموزش داده می‌شود. ویژگی های استخراجی از این تسک شامل روابط فضایی و معنایی در تصاویر است.

پاسخ سوال ۲

حل پازل

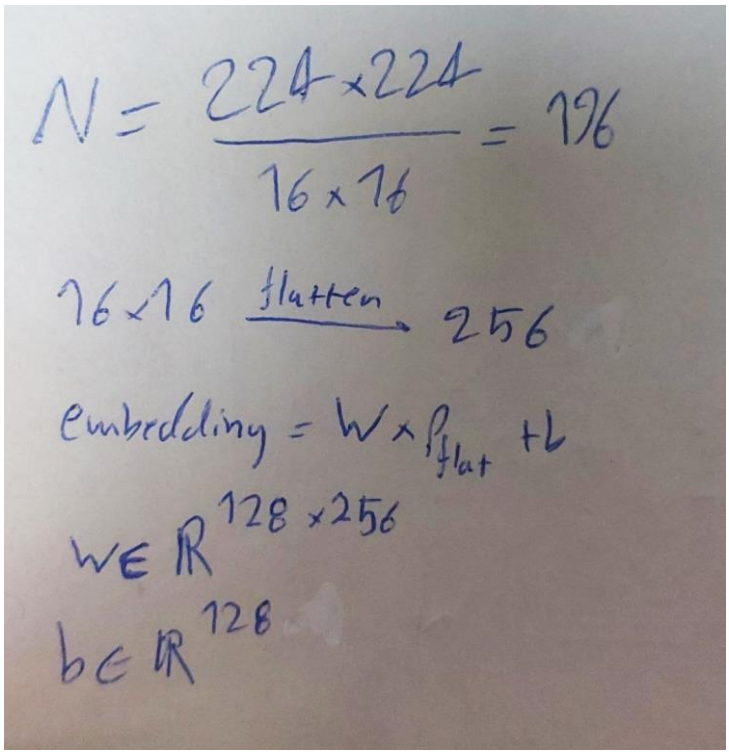
الف) تصاویر ماهواره ای قرارگیری های مشخص و پیچیده ای را دربردارند مانند خیابان ها، ساختمان ها، مزارع و غیره. با آموزش مدل بر روی وظیفه پیش‌متن حل پازل، می‌توانیم ویژگی های فضایی و ارتباط بین بخش های مختلف این تصاویر را استخراج کنیم که برای وظایف بعدی بسیار مفید واقع می‌شوند.

ب) ابتدا تصاویر ماهواره ای را به پچ های کوچک مربعی تقسیم می‌کنیم. سپس آن ها را شافل کرده و به مدل می‌دهیم تا بتواند بر اساس این پچ ها تصویر اصلی را بازسازی کند.

ج) با چرخش تصاویر ماهواره ای معمولا تغییر معنی داری میسر نمی‌شود و در نتیجه مدل ویژگی های بخصوصی را یاد نمی‌گیرد. از طرفی تصاویر ماهواره ای معمولا grayscale هستند و حتی اگر رنگی نیز باشند ویژگی های مفیدی را که می‌توان از حل پازل استخراج کرد را حاصل نمی‌کنند.

پاسخ سوال ۳

الف)



ب) $\text{embedding}[i] \leftarrow \text{embedding}[i] + \text{PE}[I] \quad , i = 1,2, \dots, 196$

بدون جاسازی موقعیتی مدل نمی‌تواند اطلاعاتی در مورد روابط فضایی بین پچ ها را یاد بگیرد. با افزودن این جاسازی به مدل این قابلیت را میدهیم که بتواند روابط نسبی یا مطلق بین پچ ها را یاد بگیرد تا بتواند ویژگی های فضایی مهم را استخراج کند تا در تسک های مختلف پردازش تصویر مفید واقع شود.

ج) با افزودن توکن [CLS] به ابتدای دنباله، تعداد توکن ها به ۱۹۷ میرسد که بعد هر توکن ۱۲۸ می باشد. در ترنسفورمر تمامی توکن ها با توکن های دیگر توسط مکانیز self-attention تعامل می‌کنند. تعامل توکن [CLS] با سایر توکن ها، باعث میشود تا این توکن تمامی اطلاعات مربوط به همه توکن ها را جمع آوری کند. این جمع آوری باعث میشود تا مدل یک دید کلی یا گلوبال به پچ ها داشته باشد، گویا کل تصویر اصلی را دیده است و ویژگی های آن را استخراج کرده است.

پاسخ سوال ۴

الف) ابتدا تصویر از انکودر ViT میگذرد تا یک embedding از تصویر بدست آید. از طرف دیگر متن مورد نظر از یک انکودر ترنسفورمر عبور میکند تا embedding آن نیز بدست آید. سپس با استفاده از cosine similiarity تشابه این دو embedding محاسبه می‌شود. مقدار حاصل نشان میدهد که چه میزان تصویر با متن همخوانی دارد.

برای تصویر «سیب قرمز» متن «یک سیب قرمز آبدار روی میز» بیشترین امتیاز را خواهد داشت. کلمات «سیب»، «قرمز» به صورت مستقیم در متن آمده اند که باعث نزدیکی بردار embedding تصویر و متن خواهد بود.

ب) این رفتار نشان میدهد که مدل به رنگ و شکل اجسام وزن بیشتری در امبدینگ میدهد. زیرا باوجود اینکه کلمه «سیب» به طور مستقیم در متن دوم آورده شده، مدل به سرخ بودن و گرد بودن اولویت بیشتری داده. همچنین نشان میدهد که واژه «سبز» نزدیکی معنایی بین تصویر و متن را کاهش میدهد که نشان دهنده بایاس مدل به رنگ، نسبت به روابط معنایی است.

پاسخ سوال ۵

الف) Attention Pooling:

- میانگین‌گیری از وزن های attention برای محاسبه feature map
- وزن های attention به صورت پویا یادگرفته میشوند و به نواحی مختلف feature map متصل میشوند که باعث میشود مدل بخش های مهم تر تصویر را یاد بگیرد.
- خروجی context-aware است و از جمع میانگین فیچر مپ ها بدست می آید.
- میتواند برای تسک های مختلف خروجی های متفاوتی بدهد.

Global Average Pooling:

- میانگین هر فیچر مپ را بر روی تمامی ابعاد محاسبه میکند. تمامی اطلاعات فضایی برای هر ویژگی در یک مقدار خلاصه می‌شود.
- خروجی یک میانگین یکنواخت از تمامی ویژگی هاست.
- از دست رفتن اطلاعات را به همراه دارد که از توانایی مدل برای تمرکز بر روی نواحی مهم می‌کاهد.

ب) ماتریس لیبل یک ماتریس $N \times N$ با درایه های ۰(نشان دهنده جفت های منفی) و ۱(نشان دهنده جفت های مثبت) است. اگر فرض کنیم که تصاویر لیبل مشترک ندارند، تنها درایه های قطر اصلی مقادیر ۱ خواهند داشت. پس ماتریس N تا ۱ و N^2-N صفر دارد.

ج) بر اساس مقاله، عملکرد zero-shot مدل کلیپ در تسک هایی ضعیف است که :

۱. نیاز به دانش تخصصی دارند مانند تصاویر پزشکی، تصاویر ماهواره‌ای و ...

۲. نیاز به نتیجه‌گیری و استنتاج های پیچیده و انتزاعی دارند مانند شمردن اشیا، تشخیص ویژگی های جزئی و ...

۳. نیاز به اطلاعات سطح بالاتری دارند مانند علائم رانندگی

دلیل ضعف کلیپ در موارد ذکر شده عبارتند از:

۱- نیاز به تعداد دیتای بالا و متنوع
۲- تکیه مدل صرفا بر الگو های بصری مانند رنگ، شکل، بافت
۳- ناتوانی زبان طبیعی در توصیف برخی تسک ها