دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

زمستان ۱۴۰۳

علیرضا میررکنی - ۴۰۱۱۰۶۶۱۷

تمرین ششم

پاسخ مسئله‌ی ۱.

یادگیری خودنظارتی (Self-Supervised Learning) یک رویکرد مدرن در یادگیری ماشین است که برای بهره‌برداری از داده‌های بدون برچسب طراحی شده است. این روش به مدل اجازه می‌دهد تا به‌صورت خودکار برچسب‌هایی را از درون داده‌ها استخراج کرده و از آن‌ها برای آموزش استفاده کند. یادگیری خودنظارتی به‌طور گسترده در حوزه‌هایی مانند بینایی کامپیوتر و پردازش زبان طبیعی به کار گرفته می‌شود.

وظایف پیش‌متن (Pretext Tasks)

وظایف پیش‌متن، وظایف مصنوعی و کمکی هستند که برای هدایت مدل به یادگیری بازنمایی‌های مفید طراحی می‌شوند. هدف از این وظایف، استخراج ویژگی‌های معنادار از داده‌های ورودی بدون نیاز به برچسب‌های دستی است. پس از آموزش با این وظایف، مدل می‌تواند در وظایف اصلی (مانند طبقه‌بندی یا شناسایی اشیا) عملکرد بهتری داشته باشد. وظایف پیش‌متن معمولاً به مدل کمک می‌کنند تا اطلاعات آماری و معنایی موجود در داده‌ها را به‌طور عمیق درک کند.

در ادامه، سه نمونه از وظایف پیش‌متن رایج به همراه ویژگی‌هایی که به مدل آموزش می‌دهند، توضیح داده شده‌اند:

الف) پیش‌بینی چرخش (Rotation Prediction)

در این وظیفه، یک تصویر ورودی به‌طور تصادفی با یکی از زوایای 0°، 90°، 180°، یا 270° چرخانده می‌شود. سپس مدل باید زاویه چرخش را پیش‌بینی کند.

ویژگی‌های آموخته‌شده:

- یادگیری ساختارهای مکانی (Spatial Structures) و هندسه اشیا.
- حساسیت به الگوهای ساختاری، خطوط و لبه‌ها برای استخراج ویژگی‌های مکانی عمیق‌تر.
- تقویت ناوردا بودن (Invariance) در برابر چرخش که در تشخیص اشیا کاربرد دارد.

ب) رنگ‌آمیزی (Colorization)

در این وظیفه، تصویر ورودی به یک تصویر سیاه‌وسفید تبدیل شده و مدل وظیفه دارد رنگ‌های اصلی تصویر را بازسازی کند. این کار نیازمند درک قوی از روابط معنایی در تصویر است.

ویژگی‌های آموخته‌شده:

- یادگیری ویژگی‌های معنایی (Semantic Features) مانند اشکال، بافت‌ها و ارتباط بین نواحی مختلف تصویر.

- استخراج وابستگی‌های محلی (Local Dependencies) بین پیکسل‌ها و درک محتوا.
- تقویت دانش مدل درباره اشیا و صحنه‌ها به کمک بازسازی رنگ.

ج) حل پازل (Jigsaw Puzzle Solving)

در این وظیفه، یک تصویر به چندین قطعه کوچک تقسیم شده و سپس این قطعات به‌طور تصادفی جابه‌جا می‌شوند. مدل باید ترتیب صحیح قطعات را بازسازی کند.

ویژگی‌های آموخته‌شده:

- یادگیری وابستگی‌های مکانی (Spatial Relationships) بین بخش‌های مختلف تصویر.
- درک ارتباط اشکال و الگوها در مناطق مجاور.
- تقویت توانایی استخراج ویژگی‌های سطح بالا و اطلاعات مکانی.

وظایف پیش‌متن در یادگیری خودنظارتی نقش کلیدی در استخراج بازنمایی‌های مؤثر از داده‌ها دارند. این وظایف به مدل کمک می‌کنند تا ویژگی‌هایی مانند وابستگی‌های مکانی، اطلاعات معنایی، و روابط بین پیکسل‌ها را بدون نیاز به برچسب‌گذاری دستی یاد بگیرد. پس از آموزش اولیه، مدل می‌تواند این ویژگی‌ها را در وظایف پایین‌دستی (مانند طبقه‌بندی یا تشخیص اشیا) به کار گیرد و عملکرد بهتری داشته باشد. یادگیری خودنظارتی به دلیل توانایی در استفاده از حجم عظیم داده‌های بدون برچسب، به یکی از رویکردهای مهم در یادگیری عمیق تبدیل شده است.

پاسخ مسئله‌ی ۲.

انتخاب وظیفه پیش‌متن: حل پازل (Jigsaw Puzzle Solving)

الف) دلایل همخوانی این وظیفه با ویژگی‌های تصاویر ماهواره‌ای:

- تصاویر ماهواره‌ای معمولاً دارای ساختارهای فضایی مشخص مانند خیابان‌ها، ساختمان‌ها و الگوهای هندسی هستند. حل پازل به مدل کمک می‌کند تا روابط فضایی (Spatial Relationships) بین اجزای تصویر را یاد بگیرد.
- این تصاویر دارای تنوع زیاد در مقیاس، جهت و شکل هستند. وظیفه حل پازل مدل را به یادگیری ویژگی‌های مکانی و ترکیب‌بندی‌های فضایی تشویق می‌کند که برای وظایفی مانند تشخیص ساختمان‌ها بسیار مفید است.
- ویژگی‌های محلی مانند لبه‌ها، بافت‌ها و خطوط در تصاویر ماهواره‌ای اهمیت زیادی دارند. حل پازل باعث استخراج چنین ویژگی‌هایی می‌شود.

ب) ابتدا هر تصویر ماهواره‌ای به چندین قطعه کوچک‌تر (مانند ۳×۳ یا ۴×۴) تقسیم می‌شود. این قطعات به‌طور تصادفی جابه‌جا شده و ترتیب اصلی آن‌ها تغییر می‌کند. سپس مدل با هدف بازسازی ترتیب صحیح قطعات آموزش می‌بیند. در طول آموزش، مدل ابتدا ویژگی‌های محلی هر قطعه را استخراج کرده و سپس روابط فضایی بین قطعات را یاد می‌گیرد. این فرآیند باعث می‌شود مدل بتواند ویژگی‌های فضایی و ساختاری معناداری را استخراج کند که برای وظایف بعدی مانند تشخیص ساختمان‌ها و طبقه‌بندی کاربری زمین مفید خواهند بود.

ج) محدودیت‌های دو وظیفه دیگر برای این نوع داده‌ها:

۱. پیش‌بینی چرخش (Rotation Prediction):

- تصاویر ماهواره‌ای اغلب از زوایای مختلف گرفته می‌شوند و جهت‌گیری آن‌ها استاندارد نیست.
- ساختمان‌ها و خیابان‌ها ممکن است به‌طور طبیعی دارای چرخش‌های مختلف باشند، بنابراین پیش‌بینی زاویه چرخش می‌تواند گمراه‌کننده باشد.
- این وظیفه برای تصاویری که جهت‌گیری مشخص دارند (مانند تصاویر اشیا از دید روبه‌رو) مناسب‌تر است، نه برای تصاویر هوایی که جهت آن‌ها متغیر است.

۲. رنگ‌آمیزی (Colorization):

- بسیاری از تصاویر ماهواره‌ای در باندهای طیفی مختلف (مانند مادون قرمز یا ترکیب‌های چندطیفی) ثبت می‌شوند که ممکن است به‌صورت سیاه‌وسفید نمایش داده نشوند.
- رنگ‌آمیزی برای بازسازی رنگ طبیعی تصاویر طراحی شده است و به اطلاعات طیفی خاص موجود در تصاویر ماهواره‌ای توجه نمی‌کند.
- در تصاویر ماهواره‌ای، اطلاعات رنگی ممکن است کمتر به ویژگی‌های معنایی مربوط باشد و بیشتر به ترکیب مواد و پوشش زمین وابسته باشد، که این کار را برای رنگ‌آمیزی دشوار می‌کند.

پاسخ مسئله‌ی ۳.

الف) تصویر ورودی دارای ابعاد ۲۲۴ × ۲۲۴ پیکسل است. هر پچ نیز دارای ابعاد ۱۶ × ۱۶ پیکسل می‌باشد.
ابتدا تعداد پچ‌ها را محاسبه می‌کنیم:

$$\frac{۲۲۴}{۱۶} \times \frac{۲۲۴}{۱۶} = ۱۴ \times ۱۴ = ۱۹۶$$

بنابراین، تصویر به ۱۹۶ پچ تقسیم می‌شود.
هر پچ دارای ۲۵۶ = ۱۶ × ۱۶ پیکسل است. این مقدار به یک بردار ویژگی مسطح‌شده با طول ۲۵۶ تبدیل می‌شود. سپس، با استفاده از یک لایه خطی (Linear) این بردار به یک بردار جاسازی شده با ابعاد ۱۲۸ تبدیل می‌شود:

$$Z = XW + b$$

که در آن:

- X ماتریس ورودی با ابعاد ۲۵۶ × ۱۹۶ است.
- W ماتریس وزن با ابعاد ۱۲۸ × ۲۵۶ است.
- b بردار بایاس با ابعاد ۱۲۸ است.

خروجی Z دارای ابعاد ۱۲۸ × ۱۹۶ خواهد بود.

ب) از آنجایی که مدل ViT ارتباطات فضایی بین پچ‌ها را به‌صورت صریح ذخیره نمی‌کند، باید اطلاعات مربوط به موقعیت نسبی پچ‌ها به مدل اضافه شود. این کار با اضافه کردن بردارهای جاسازی موقعیت (Positional Encoding) انجام می‌شود. برای این منظور، یک بردار موقعیت با ابعاد مشابه بردارهای جاسازی شده (۱۲۸) برای هر پچ تعریف می‌شود. سپس این بردار به بردار ویژگی هر پچ اضافه می‌شود:

$$Z' = Z + P$$

که در آن:

- Z بردارهای ویژگی پچ‌ها با ابعاد ۱۲۸ × ۱۹۶ است.
- P بردارهای جاسازی موقعیت با ابعاد ۱۲۸ × ۱۹۶ است.

اهمیت جاسازی موقعیت را می‌توان در موارد زیر خلاصه کرد:

- حفظ اطلاعات فضایی بین پچ‌ها برای درک ساختار تصویر.
- جلوگیری از ازدست‌رفتن نظم توالی پچ‌ها که برای یادگیری روابط مکانی ضروری است.
- بهبود عملکرد مدل در استخراج ویژگی‌های محلی و روابط بین اجزا.

ج) توکن [CLS] یک بردار ویژه است که در ابتدای دنباله ورودی قرار می‌گیرد. این بردار به‌عنوان نماینده کل دنباله استفاده می‌شود و در طول فرآیند توجه (Attention) به‌روزرسانی می‌شود.

ابعاد توکن [CLS] مشابه سایر بردارهای جاسازی شده برابر ۱۲۸ است. نقش‌های این توکن را می‌توان در موارد زیر خلاصه کرد:

- خلاصه‌سازی اطلاعات کلی تصویر پس از فرآیند توجه در لایه‌های ترنسفورمر.
- استفاده به‌عنوان ورودی به لایه نهایی برای انجام وظایف پایین‌دستی مانند طبقه‌بندی یا تشخیص اشیا.
- ساده‌سازی پردازش خروجی مدل با تمرکز روی یک بردار نماینده به‌جای نیاز به پردازش تمام بردارهای پچ‌ها.

در نهایت، بردار [CLS] به لایه‌های بعدی مانند یک طبقه‌بند خطی داده می‌شود تا وظایف خروجی مانند پیش‌بینی برچسب کلاس را انجام دهد.

در مدل ViT، تصویر ورودی به پچ‌های کوچک تقسیم شده و هر پچ به یک بردار جاسازی شده تبدیل می‌شود. سپس با اضافه کردن اطلاعات موقعیت مکانی، مدل قادر به یادگیری ارتباطات فضایی می‌شود. توکن ویژه

پاسخ مسئله‌ی ۴.

الف) مدل CLIP برای محاسبه شباهت بین تصویر و متون توصیفی از یک فضای مشترک تعبیه (Embedding Space) استفاده می‌کند. این فرآیند شامل مراحل زیر است:

- ابتدا، تصویر و متن‌ها به ترتیب توسط یک شبکه عصبی بینایی و یک شبکه زبانی کدگذاری می‌شوند و به بردارهای ویژگی (Feature Vectors) در فضای تعبیه تبدیل می‌شوند.
- سپس، شباهت بین بردارهای تصویر (v_i) و متن (t_j) با استفاده از ضرب داخلی کسینوسی محاسبه می‌شود:

$$S(i,j) = \frac{v_i \cdot t_j}{\|v_i\|\|t_j\|} \tag{۱}$$

که در آن $S(i,j)$ میزان شباهت بین تصویر i و متن j را نشان می‌دهد.

احتمالاً متن اول («یک سیب قرمز آبدار روی میز») بالاترین امتیاز را می‌گیرد زیرا:

- رنگ «قرمز» در هر دو (تصویر و متن) وجود دارد که یک ویژگی بصری کلیدی است.
- شیء «سیب» دقیقاً با محتوای تصویر مطابقت دارد.
- موقعیت مکانی (روی میز) به‌طور ضمنی می‌تواند با مدل تطبیق داشته باشد، اما مهم‌تر از همه، ویژگی‌های بصری کلی (رنگ و نوع شیء) برتری دارند.

ب) اگر مدل، متن سوم («یک توپ قرمز درخشان») را بالاتر از متن دوم («یک سیب سبز») رتبه‌بندی کند، این رفتار نشان می‌دهد که فضای تعبیه (Embedding Space) یادگرفته‌شده توسط CLIP بیشتر به ویژگی‌های سطح پایین (Low-level Features) مانند رنگ (قرمز) حساس است تا ویژگی‌های سطح بالا (High-level Features) مانند نوع شیء (سیب).

این مسئله می‌تواند ناشی از همبستگی قوی بین رنگ (قرمز) و ویژگی‌های بصری در داده‌های آموزشی مدل باشد که باعث شده مدل به جای تمرکز بر معنای شیء، به شباهت‌های ظاهری اولویت بدهد.

همچنین این رفتار نشان می‌دهد که بردارهای ویژگی متون و تصاویر در فضای تعبیه ممکن است به‌طور ناهماهنگ خوشه‌بندی شده باشند، به‌طوری‌که مفاهیم مرتبط با رنگ در فاصله کمتری نسبت به مفاهیم معنایی قرار گرفته‌اند.

این مسئله ضعف مدل در یادگیری بازنمایی‌های غنی معنایی را نشان می‌دهد و ممکن است نیاز به تنظیم دقیق (Fine-tuning) با داده‌های متنوع‌تر برای تعادل بهتر بین ویژگی‌های سطح پایین و بالا داشته باشد.

پاسخ مسئله‌ی ۵.

الف) Global Average Pooling (GAP) و Attention Pooling دو روش برای کاهش ابعاد ویژگی‌ها و ترکیب اطلاعات فضایی در شبکه‌های عصبی هستند. در این بخش، این دو روش با جزئیات بیشتری مقایسه شده‌اند:

۱. Global Average Pooling (GAP):

- GAP ویژگی‌های مکانی را با محاسبه میانگین مقدار هر کانال در سراسر فضا (Height × Width) کاهش می‌دهد.
- خروجی یک بردار ویژگی با ابعاد ثابت است که به‌طور مؤثری پیچیدگی شبکه را کاهش می‌دهد.
- معادله ریاضی برای این عملیات به‌صورت زیر است:

$$GAP(x) = \frac{۱}{H \times W} \sum_{i=۱}^{H} \sum_{j=۱}^{W} x_{i,j}$$

که در آن H و W ابعاد فضایی هستند و $x_{i,j}$ مقدار ویژگی در مکان (i, j) است.

- این روش ساده و محاسباتی کارآمد است اما اطلاعات مکانی دقیق و روابط معنایی محلی را حذف می‌کند.

۲. Attention Pooling:

- این روش به‌صورت پویا وزن‌هایی را برای بخش‌های مختلف تصویر بر اساس اهمیت آن‌ها محاسبه می‌کند.
- وزن‌ها با استفاده از یک مکانیزم توجه محاسبه می‌شوند که اطلاعات مهم‌تر را برجسته می‌کند.
- معادله ریاضی این روش با استفاده از ماتریس توجه به شکل زیر است:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

$$Z = AV$$

که در آن Q، K، و V به ترتیب بردارهای پرسش (Query)، کلید (Key) و مقدار (Value) هستند و d_k بعد کلیدها است.

- Attention Pooling می‌تواند اطلاعات فضایی و معنایی بیشتری را حفظ کند و در وظایف پیچیده‌تر عملکرد بهتری دارد، اما نیازمند محاسبات بیشتری است.

ب) در یادگیری تضادی، هدف به حداکثر رساندن شباهت بین جفت‌های مرتبط (مثلاً تصویر و متن منطبق) و حداقل رساندن شباهت بین جفت‌های نامرتبط است. این فرایند با استفاده از ماتریس لیبل $N \times N$ تعریف می‌شود:

- N تعداد نمونه‌ها در یک مینی‌بچ است.
- ماتریس L شامل درایه‌هایی به‌صورت زیر است:

$$L_{i,j} = \begin{cases} ۱ & i = j \\ ۰ & \text{otherwise} \end{cases}$$

- درایه‌های ۱ بیانگر جفت‌های مثبت (مرتبط) هستند و درایه‌های ۰ نشان‌دهنده جفت‌های منفی (نامرتبط) می‌باشند.

در نتیجه تعداد درایه‌های ناصفر برابر $N^۲ - N$ می‌باشد. این مقدار نشان‌دهنده جفت‌های نامتطابق است که مدل باید برای تمایز آن‌ها یادگیری انجام دهد.

ج) مدل CLIP در وظایف خاص مانند طبقه‌بندی تصاویر پیچیده، تصاویر پزشکی یا تشخیص اشیاء کوچک ضعیف‌تر عمل می‌کند. همچنین در وظایفی که نیاز به استنتاج عمیق معنایی دارند، دچار افت دقت می‌شود.

علت این ضعف را می‌توان در موراد زیر خلاصه کرد:

- **تعصب به ویژگی‌های سطح پایین:** مدل عمدتاً بر ویژگی‌های ساده مانند رنگ، بافت و اشکال تمرکز دارد و ممکن است ویژگی‌های معنایی سطح بالا را به‌خوبی ثبت نکند.
- **کمبود داده‌های آموزشی:** داده‌های آموزشی ممکن است شامل تصاویر خاص (مانند تصاویر ماهواره‌ای یا پزشکی) نبوده باشند که موجب ضعف تعمیم‌پذیری مدل شده است.
- **عدم تطبیق دقیق فضای تعبیه:** فضای تعبیه یادگرفته‌شده ممکن است برای تطابق جفت‌های پیچیده، اطلاعات معنایی کافی را حفظ نکند.
- **چالش وظایف انتزاعی:** وظایفی مانند شمارش اشیاء یا توصیف ارتباطات پیچیده بین اشیاء برای مدل دشوارتر است.

بنابراین می‌توان نتیجه گرفت که مدل CLIP در تنظیمات Zero-shot عملکرد قوی دارد، اما در مواجهه با داده‌های خاص و وظایف پیچیده که به اطلاعات معنایی عمیق نیاز دارند، محدودیت‌هایی نشان می‌دهد. این امر نشان‌دهنده نیاز به تنظیم دقیق و داده‌های متنوع‌تر برای بهبود عملکرد است.