نام و نام خانوادگی:

رضا قربانی پاجی

شماره دانشجویی:

403206565

تمرین ششم درس یادگیری ماشین

سوال 1

مفهوم یادگیری خودنظارتی (Self-Supervised Learning) و وظایف پیش‌متن (Pretext Task)

یادگیری خودنظارتی نوعی روش یادگیری است که در آن مدل بدون نیاز به برچسب‌های دستی داده‌ها آموزش می‌بینند. این روش به گونه‌ای طراحی شده است که مدل از داده‌های بدون برچسب برای تولید سیگنال‌های آموزشی استفاده می‌کند و این کار معمولاً با ایجاد وظایف پیش‌متن (pretext task) انجام می‌شود. وظایف پیش‌متن اهداف مصنوعی یا کمکی هستند که برای مدل تعریف می‌شوند تا ویژگی‌های مفید داده‌ها را بیاموزد.

وظایف پیش‌متن معمولاً بر اساس دستکاری داده‌ها به نحوی تعریف می‌شوند که مدل بتواند ساختار داده‌ها یا روابط داخلی آن‌ها را درک کند. پس از انجام این وظایف، مدل برای وظایف اصلی (مانند دسته‌بندی، تشخیص اشیا، یا ترجمه) به کار گرفته می‌شود.

الف) پیش‌بینی چرخش(Rotation Prediction)

در این وظیفه، تصویر اصلی به چندین حالت چرخش مختلف (مثلاً 0، 90، 180، و 270 درجه) تبدیل می‌شود. مدل باید یاد بگیرد که چرخش تصویر را تشخیص دهد. هدف این است که مدل ویژگی‌های فضایی (spatial features)و ساختار اشیاء در تصویر را یاد بگیرد.

- **ویژگی‌های آموزش داده‌شده :**مدل یاد می‌گیرد تا ساختار شیء و نحوه قرارگیری آن در تصویر را شناسایی کند. این ویژگی‌ها می‌توانند در وظایفی مانند دسته‌بندی تصویر مفید باشند.

ب) رنگ‌آمیزی(Colorization)

در این وظیفه، یک تصویر رنگی به تصویر خاکستری (grayscale) تبدیل می‌شود و مدل باید تصویر خاکستری را به تصویر رنگی تبدیل کند. هدف از این وظیفه این است که مدل روابط بین کانال‌های رنگی مختلف و همچنین ساختار و بافت اشیاء را یاد بگیرد.

- **ویژگی‌های آموزش داده‌شده :**مدل توانایی درک روابط بین پیکسل‌ها و کانال‌های رنگی را پیدا می‌کند و همچنین قادر به تشخیص الگوهای متنی و بافت در تصویر می‌شود.

ج) حل پازل(Jigsaw Puzzle Solving)

در این وظیفه، یک تصویر به چند قطعه تقسیم می‌شود و این قطعات به طور تصادفی جابجا می‌شوند. مدل باید یاد بگیرد که قطعات را به ترتیب درست خود بچیند. این وظیفه به مدل کمک می‌کند تا روابط مکانی (spatial relationships)و انسجام ساختاری تصویر را درک کند.

- **ویژگی‌های آموزش داده‌شده :**مدل یاد می‌گیرد که بافت، لبه‌ها، و ترتیب منطقی قطعات تصویر را تحلیل کند و روابط فضایی بین بخش‌های مختلف تصویر را شناسایی کند.

سوال 2

الف) چرا این وظیفه با ساختار و ویژگی‌های تصاویر ماهواره‌ای همخوانی دارد؟

وظیفه انتخاب‌شده: پیش‌بینی چرخش(Rotation Prediction)

- تصاویر ماهواره‌ای دارای ویژگی‌های مکانی و فضایی مشخصی هستند، مانند الگوهای تکرارشونده ساختمان‌ها، جاده‌ها، مناطق کشاورزی و غیره. پیش‌بینی چرخش به مدل کمک می‌کند تا ساختارهای فضایی را بهتر شناسایی کرده و روابط بین اشیاء در تصویر را درک کند.
- این وظیفه با ساختار تصاویر ماهواره‌ای همخوانی دارد، زیرا تصاویر ماهواره‌ای ممکن است در جهت‌های مختلف ثبت شده باشند و مدل نیاز دارد تا بدون وابستگی به جهت‌گیری اولیه، ویژگی‌های اصلی تصویر را شناسایی کند.
- همچنین، این وظیفه کمک می‌کند که مدل به‌جای یادگیری الگوهای خاص جهت‌گیری، بر یادگیری روابط و الگوهای کلی تمرکز کند.

ب) چگونه می‌توان این وظیفه پیش‌متن را روی این داده‌ها اعمال کرد؟

برای اعمال وظیفه پیش‌بینی چرخش بر روی تصاویر ماهواره‌ای، مراحل زیر انجام می‌شود:

1. **ایجاد داده‌های چرخشی:**
   - تصاویر ماهواره‌ای اصلی را به صورت تصادفی به زاویه‌های 0، 90، 180 و 270 درجه بچرخانید.
   - به هر تصویر چرخیده یک برچسب عددی تخصیص دهید که نشان‌دهنده زاویه چرخش آن باشد (مثلاً: 0 برای چرخش 0 درجه، 1 برای 90 درجه، و ...).
2. **آموزش مدل:**
   - مدل را آموزش دهید تا بر اساس ساختارهای مکانی موجود در تصویر، زاویه چرخش را پیش‌بینی کند.
   - در طول فرآیند آموزش، مدل به ویژگی‌های مکانی و ساختاری تصویر حساس می‌شود و آن‌ها را یاد می‌گیرد.

3. **استفاده از ویژگی‌های آموخته‌شده:**

- پس از اتمام آموزش، وزن‌ها و ویژگی‌های استخراج‌شده از مدل را به وظایف بعدی (مانند تشخیص ساختمان‌ها یا طبقه‌بندی کاربری زمین) منتقل کنید.

ج) دو وظیفه دیگر چه محدودیت‌هایی برای این نوع داده‌ها دارند؟

1. **رنگ‌آمیزی:(Colorization)**

- **محدودیت ۱ :**تصاویر ماهواره‌ای معمولاً شامل باندهای چندطیفی یا مادون قرمز هستند که رنگ‌آمیزی آن‌ها می‌تواند پیچیده باشد و ممکن است اطلاعات ارزشمند طیفی از دست برود.
- **محدودیت ۲ :**این وظیفه بیشتر برای تصاویر RGB مؤثر است و ممکن است برای تصاویر ماهواره‌ای که اطلاعات مکانی و طیفی در آن‌ها اهمیت بیشتری دارد، مناسب نباشد.

2. **حل پازل:(Jigsaw Puzzle Solving)**

- **محدودیت ۱ :**تصاویر ماهواره‌ای ممکن است شامل الگوهای تکراری (مانند مزارع یا مناطق جنگلی) باشند که تشخیص قطعات و مرتب‌سازی آن‌ها را برای مدل دشوار می‌کند.
- **محدودیت ۲ :**این وظیفه بیشتر بر ویژگی‌های محلی تمرکز دارد و ممکن است روابط بزرگ‌مقیاس بین اشیاء (مانند جاده‌ها و ساختمان‌ها) را در نظر نگیرد.

سوال 3

پاسخ سؤال ۳:

الف) محاسبه تعداد کل پچ‌ها (N) و توضیح فرآیند تبدیل خطی به ابعاد ۱۲۸:

1. **محاسبه تعداد پچ‌ها:(N)**
   - تصویر ورودی دارای ابعاد 224×224 پیکسل است.
   - تصویر به پچ‌هایی با ابعاد 16×16 تقسیم می‌شود.
   - تعداد پچ‌ها N برابر است با تعداد بخش‌هایی که تصویر به آن‌ها تقسیم می‌شود :

$$N = \frac{224}{16} \times \frac{224}{224} = 14 \times 14 = 196$$

   - بنابراین، تصویر به 196 پچ تقسیم می‌شود.
2. **تبدیل خطی هر پچ به یک بردار ۱۲۸ بعدی:**
   - هر پچ 16×16 شامل 256 مقدار پیکسل است (تعداد کل پیکسل‌های هر پچ).
   - ابتدا هر پچ به یک بردار خطی 256 بعدی فلت می‌شود.
   - سپس از یک لایه خطی (Linear Layer) استفاده می‌شود تا این بردار 256 بعدی به یک بردار 128 بعدی نگاشت شود

$$h = W.x + b$$

که در آن :

- x بردار ورودی 256 بعدی است.
- W ماتریس وزن با ابعاد 256×128 است.
- b بایاس با ابعاد 128است.

   - این عملیات باعث کاهش ابعاد به 128 می‌شود و ویژگی‌های فشرده و قابل استفاده برای ورودی مدل ایجاد می‌کند.

ب) جاسازی موقعیتی (Positional Embedding) و دلیل اهمیت آن:

1. **جاسازی موقعیتی چیست؟**

- در مدل‌های ترانسفورمر، اطلاعات موقعیت نسبی یا ترتیبی داده‌ها در ورودی حفظ نمی‌شود.
- برای رفع این مشکل، به هر پچ ورودی یک بردار موقعیتی (Positional Embedding) اضافه می‌شود که نشان‌دهنده مکان آن پچ در تصویر است.
- بردار موقعیتی یک بردار عددی است که به بردار ویژگی هر پچ اضافه می‌شود تا مدل بتواند موقعیت نسبی پچ‌ها را درک کند.

2. **نحوه اضافه کردن جاسازی موقعیتی:**

- پس از تبدیل هر پچ به یک بردار 128 بعدی، یک بردار جاسازی موقعیتی 128 بعدی به آن اضافه می‌شود.
- اگر $N = 196$ تعداد کل پچ‌ها باشد، یک ماتریس جاسازی موقعیتی با ابعاد 128×196ایجاد می‌شود و به ویژگی‌های پچ‌ها اضافه می‌گردد.

3. **اهمیت:**

- این عملیات به مدل کمک می‌کند تا ساختار فضایی تصویر را حفظ کند و روابط بین پچ‌ها را درک نماید.
- بدون این اطلاعات، مدل نمی‌تواند ترتیب مکانی پچ‌ها را در تصویر شناسایی کند و ممکن است عملکرد آن کاهش یابد.

ج) ساخت توکن ویژه [CLS] و نقش آن:

1. **ساخت توکن ویژه:[CLS]**

- توکن [CLS] یک بردار ویژگی اضافی است که به ورودی مدل اضافه می‌شود.
- این توکن معمولاً به صورت یک بردار عددی با ابعاد مشابه سایر بردارهای ورودی (یعنی 128 بعدی) تعریف می‌شود.

- قبل از ارسال پچ‌ها به مدل، این توکن در ابتدای توالی پچ‌ها قرار می‌گیرد.

2. **نقش توکن:[CLS]**

- این توکن به عنوان نماینده کل تصویر عمل می‌کند.
- در پایان فرآیند پردازش توسط مدل، بردار ویژگی مربوط به [CLS] برای وظایف پایین‌دستی مانند دسته‌بندی تصویر یا تشخیص شیء استفاده می‌شود.
- به عبارت دیگر، مدل اطلاعات کل تصویر را در توکن [CLS] فشرده‌سازی می‌کند.

3. **ابعاد ورودی نهایی:**

- تعداد کل پچ‌ها N = 196 است و یک توکن [CLS] اضافه می‌شود

$$N_{final} = 196 + 1 = 197$$

- بنابراین ورودی نهایی مدل دارای ابعاد 128×197 است.

سوال 4

الف) نحوه محاسبه شباهت توسط CLIP :

1. **پردازش تصویر:**
   - تصویر «سیب قرمز» از طریق یک شبکه عصبی) مانند ResNet یا (Vision Transformer عبور داده می‌شود و به یک بردار ویژگی (embedding) در فضای نمایش تبدیل می‌شود.
2. **پردازش متن:**
   - هر یک از جملات متنی («یک سیب قرمز روی میز»، «یک سیب سبز آویزان از درخت»، «یک توپ قرمز درخشان») از طریق یک مدل زبان) مانند (Transformer پردازش شده و به بردارهایی در همان فضای نمایش تبدیل می‌شوند.
3. **محاسبه شباهت:**
   - شباهت بین بردار تصویر و بردار هر متن با استفاده از **ضرب داخلی (dot product)** یا **کسینوس شباهت (cosine similarity)** محاسبه می‌شود. مقادیر بالاتر نشان‌دهنده شباهت بیشتر هستند.

- به احتمال زیاد، جفت تصویر و جمله «یک سیب قرمز روی میز» بالاترین امتیاز شباهت را خواهد داشت.
- **دلیل :**این متن دقیق‌ترین توصیف از ویژگی‌های بصری تصویر (رنگ قرمز و سیب بودن) است و ویژگی‌های متنی و تصویری در فضای نمایش نزدیک به هم قرار می‌گیرند.

ب)

رفتار مدل:

- اگر متن «یک توپ قرمز درخشان» رتبه بالاتری از «یک سیب سبز» بگیرد، نشان می‌دهد که مدل CLIP ویژگی «رنگ قرمز» را در فضای نمایش برجسته‌تر از نوع شیء (سیب یا توپ) در نظر گرفته است.
- مدل CLIP در فضای نمایش خود، **رنگ قرمز** را به عنوان ویژگی غالب شناسایی کرده است. بنابراین، بردار متن «یک توپ قرمز درخشان» از نظر مدل، به بردار تصویر «سیب قرمز» نزدیک‌تر است، زیرا هر دو ویژگی اصلی رنگ قرمز را به اشتراک می‌گذارند.
- این رفتار نشان می‌دهد که فضای نمایش یادگرفته‌شده توسط CLIP ، ویژگی‌های بصری (مانند رنگ، بافت) را به شکلی سازمان‌دهی کرده که روابط مشترک میان تصویر و متن تقویت شوند.

سوال 5

لف) مقایسه مکانیزم Attention Pooling با Global Average Pooling از نظر عملکرد و تولید خروجی

1. **Global Average Pooling (GAP):**
   - **نحوه کار :**در این روش، برای هر کانال ویژگی، مقادیر پیکسل‌ها در کل تصویر به طور میانگین گرفته می‌شود. این عملیات یک بردار تک‌بعدی تولید می‌کند که نشان‌دهنده اطلاعات کلی تصویر است.
   - **عملکرد:**
     - GAPسریع و محاسباتی سبک است.
     - این روش اطلاعات فضایی (Spatial Information) تصویر را از بین می‌برد و تنها ویژگی‌های کلی (Global Features) را حفظ می‌کند.
   - **مزایا:**
     - ساده و کم‌هزینه.
     - برای مدل‌هایی که نیاز به خلاصه‌سازی کلی دارند، مناسب است.
   - **معایب:**
     - اطلاعات دقیق موقعیت مکانی اشیاء را حذف می‌کند.
2. **Attention Pooling:**
   - **نحوه کار :**در این روش، وزن‌های توجه (Attention Weights) برای هر موقعیت یا کانال محاسبه می‌شوند. سپس خروجی نهایی بر اساس این وزن‌ها به صورت ترکیب خطی از ویژگی‌های ورودی تولید می‌شود.
   - **عملکرد:**
     - این روش اطلاعات فضایی و وزنی را حفظ می‌کند و ویژگی‌های مهم‌تر را برجسته می‌سازد.
     - می‌تواند به تصاویر با اطلاعات پیچیده توجه کند.
   - **مزایا:**
     - عملکرد بهتری در استخراج ویژگی‌های محلی و مهم‌تر.
     - حفظ جزئیات فضایی.
   - **معایب:**

- محاسبات سنگین‌تر نسبت به GAP.
- نیاز به حافظه و زمان بیشتر.

مقایسه کلی:

- GAP ساده‌تر و سریع‌تر است اما اطلاعات فضایی را از بین می‌برد.
- Attention Pooling پیچیده‌تر و قدرتمندتر است و اطلاعات بیشتری درباره ساختار تصویر حفظ می‌کند.

ب) تعداد درایه‌های صفر در ماتریس بر اساس یادگیری Contrastive

- **ماتریس لیبل N×N:**
  - این ماتریس شامل مقادیر صفر و یک است. درایه i,j در این ماتریس برابر است با:
    - 1، اگر تصویر i و تصویر j متعلق به یک کلاس باشند.
    - 0، اگر تصویر i و تصویر j متعلق به کلاس‌های متفاوت باشند.
  - این ماتریس برای آموزش مدل با یادگیری Contrastive استفاده می‌شود.

- **تعداد درایه‌های صفر:**
  - اگر تعداد کلاس‌ها بسیار زیاد باشد و تصاویر از کلاس‌های متفاوت باشند، اکثر درایه‌های این ماتریس صفر خواهند بود.
  - به طور کلی، تعداد درایه‌های صفر در این ماتریس برابر است با: N^2 تعداد جفت‌های تصاویر با کلاس مشابه
  - اگر k تعداد تصاویر در یک کلاس باشد، تعداد جفت‌های مشابه برابر است با:

$$\sum_{c} \binom{k_c}{2}$$

ج) قدرت مدل CLIP در تسک‌های Zero-shot و ضعف آن

1. **قدرت مدل CLIP در تسک‌های Zero-shot:**
   - CLIP قادر است بدون آموزش مستقیم روی تسک خاص، عملکرد خوبی ارائه دهد.
   - این مدل از فضای نمایش مشترک تصویر و متن استفاده می‌کند و می‌تواند توصیفات متنی را برای تصاویر ناشناخته تفسیر کند.
   - کاربرد آن در تسک‌هایی مانند دسته‌بندی تصاویر یا جستجوی متنی بسیار مؤثر است

2. **ضعف در برخی تسک‌ها:**
   - CLIPممکن است در تسک‌هایی که نیازمند اطلاعات بسیار دقیق مکانی یا جزئی هستند، ضعیف عمل کند.
   - در تسک‌هایی که داده‌های مورد نظر خارج از دامنه یادگیری مدل هستند (مانند داده‌های تخصصی یا بسیار پیچیده)، عملکرد مدل کاهش می‌یابد.
   - دلیل این ضعف، وابستگی مدل به کیفیت و گستردگی داده‌های پیش‌پردازش‌شده است. اگر داده‌های آموزشی دامنه خاصی را پوشش ندهند، مدل توانایی تعمیم در آن دامنه را ندارد.