به نام خدا

علی قاسم زاده 401106339

تمرین 6 ام یادگیری ماشین

1. مفهوم وظیفه‌ی پیش‌متن (Pretext Task) در یادگیری خودنظارتی:

وظیفه پیش‌متن یک وظیفه‌ی مصنوعی و ساده است که برای تولید برچسب‌های شبه‌نظارتی از داده‌های بدون برچسب طراحی می‌شود. این وظیفه کمک می‌کند مدل ویژگی‌های مفیدی از داده‌ها استخراج کند که بعداً می‌توان در وظایف اصلی (مثل طبقه‌بندی یا تشخیص اشیا) از آن‌ها استفاده کرد.

وظایف پیش‌متن در یادگیری خودنظارتی به مدل کمک می‌کنند تا نمایش‌های (Representations) غنی از داده‌ها را بدون نیاز به برچسب‌گذاری دستی بیاموزد. این روش‌ها عموماً در یادگیری ویژگی‌های تصویر، ویدیو یا متن استفاده می‌شوند.

توضیح سه وظیفه‌ی پیش‌متن:

الف) پیش‌بینی چرخش:(Rotation Prediction)

در این وظیفه، یک تصویر با زاویه‌های مختلف (مثل 0، 90، 180، و 270 درجه) چرخانده می‌شود. مدل باید یاد بگیرد که زاویه چرخش تصویر را پیش‌بینی کند.

- **نوع ویژگی‌های آموخته‌شده :**ویژگی‌های مرتبط با ساختار هندسی تصویر.
- **کاربرد :**یادگیری درک چیدمان و شکل اشیا در تصاویر.

ب) رنگ‌آمیزی:(Colorization)

در این وظیفه، تصویر رنگی به‌صورت سیاه‌وسفید تبدیل می‌شود و مدل باید یاد بگیرد که تصویر را به حالت رنگی اصلی بازگرداند.

- **نوع ویژگی‌های آموخته‌شده :**ویژگی‌های رنگ و بافت در تصویر.
- **کاربرد :**یادگیری ویژگی‌های سطح بالا برای درک اطلاعات رنگی در تصاویر.

ج) حل پازل:(Jigsaw Puzzle Solving)

در این وظیفه، تصویر به چند قطعه تقسیم شده و این قطعات به‌صورت تصادفی جابه‌جا می‌شوند. مدل باید یاد بگیرد که قطعات را در ترتیب اصلی‌شان بازسازی کند.

- **نوع ویژگی‌های آموخته‌شده :**روابط فضایی و زمینه‌ای بین قسمت‌های مختلف تصویر.
- **کاربرد :**تقویت توانایی درک ارتباطات میان‌قسمتی در تصاویر.

.2

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

1. **ساختار فضایی پیچیده :**تصاویر ماهواره‌ای اغلب شامل بخش‌های مختلفی مثل ساختمان‌ها، جاده‌ها، درختان و زمین‌های کشاورزی هستند که ارتباط فضایی و زمینه‌ای بین این بخش‌ها اهمیت زیادی داره. حل پازل می‌تونه این روابط رو یاد بگیره.
2. **ثبات در بافت و الگوها :**در تصاویر ماهواره‌ای، الگوهای تکرارشونده مثل ردیف‌های ساختمان یا مزارع وجود داره که این وظیفه می‌تونه اون‌ها رو استخراج کنه.

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

1. **تقسیم تصاویر :**هر تصویر ماهواره‌ای رو به چندین قطعه (مثلاً 3×3 یا 4×4) تقسیم می‌کنیم.
2. **جابه‌جایی قطعات :**ترتیب قطعات رو به‌صورت تصادفی تغییر می‌دیم.
3. **آموزش مدل :**مدل باید یاد بگیره که قطعات رو به ترتیب اصلی‌شون بازگردونه.
4. **نتیجه :**مدل به‌طور خودکار روابط فضایی و ساختاری بین بخش‌های تصویر رو یاد می‌گیره.

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

1. **پیش‌بینی چرخش:(Rotation Prediction)**

- **محدودیت :**در تصاویر ماهواره‌ای، چرخش‌های 90، 180 یا 270 درجه ممکنه کمتر معنا داشته باشن، چون دید از بالا ثبات بیشتری داره. مثلاً چرخوندن تصویر مزرعه یا شهر ممکنه اطلاعات مفیدی اضافه نکنه.

2. **رنگ‌آمیزی:(Colorization)**

- **محدودیت :**تصاویر ماهواره‌ای اغلب در فرمت‌های چند طیفی یا سیاه‌وسفید استفاده می‌شن. این وظیفه ممکنه نتونه اطلاعات خاصی مثل روابط فضایی یا الگوهای کاربردی زمین رو به‌خوبی استخراج کنه.
- همچنین رنگ‌آمیزی برای تصاویر چند طیفی (مثل باندهای مادون قرمز) غیرعملی خواهد بود.

.3

الف) تعداد کل پچ‌ها (N) و فرآیند تبدیل هر پچ به جاسازی ۱۲۸ بعدی

1. **تعداد کل پچ‌ها:(N)**

تصویر ورودی با ابعاد به پچ‌هایی با ابعاد 16×16 تقسیم می‌شود:

$$N=(224/16)\times(224/16)=14\times14=196$$

بنابراین تعداد کل پچ‌ها **۱۹۶** است.

2. **فرآیند تبدیل پچ به جاسازی ۱۲۸ بعدی:**

- هر پچ شامل $16\times16=256$ مقدار پیکسل است که به یک بردار 256 بعدی تبدیل می‌شود.
- سپس از یک ماتریس خطی W با ابعاد 256×128 برای تبدیل بردار 256 بعدی به بردار 128 بعدی استفاده می‌شود.
  این تبدیل به صورت زیر انجام می‌شود:

$$Embedding=W\cdot x+b$$

که x بردار پچ ورودی، W ماتریس وزن و b بردار بایاس است.

ب) اضافه کردن جاسازی موقعیتی:(Positional Embedding)

1. **چگونگی اضافه شدن جاسازی موقعیتی:**

- به هر جاسازی پچ (128 بعدی) یک بردار موقعیت (128 بعدی) اضافه می‌شود.
- اگر E_i جاسازی پچ i ام و P_i بردار موقعیت i ام باشد، جاسازی نهایی برابر است با:

$$E_i' = E_i + P_i$$

2. **اهمیت جاسازی موقعیتی:**

- مدل‌های ترنسفورمر توالی ورودی را بدون توجه به ترتیب مکانی پردازش می‌کنند.
- جاسازی موقعیتی اطلاعات فضایی و مکانی را به مدل اضافه می‌کند و باعث یادگیری روابط فضایی بین پچ‌ها می‌شود.

ج) ساخت دنباله ورودی و نقش توکن ویژه:[CLS]

1. **ساخت دنباله ورودی:**

- به دنباله جاسازی پچ‌ها یک توکن ویژه ([CLS]) اضافه می‌شود که در ابتدای دنباله قرار می‌گیرد.
- ابعاد دنباله ورودی:

  تعداد بردار ها = N + 1 = 197

  هر بردار دارای 128 بعد است.
  بنابراین ابعاد نهایی دنباله ورودی : 197× 128 است.

2. **نقش و کاربرد توکن ویژه:[CLS]**

- توکن [CLS] به‌عنوان نماینده کل دنباله عمل می‌کند.
- پس از پردازش در رمزگذار، بردار نهایی [CLS] اطلاعات کل دنباله را خلاصه کرده و معمولاً برای وظایفی مثل طبقه‌بندی استفاده می‌شود.
- این بردار به یک لایه خطی متصل شده و برای پیش‌بینی استفاده می‌شود.

.4

الف) نحوه محاسبه شباهت در مدل:CLIP

1. **تبدیل تصویر و متن به بردارهای جاسازی:**

- تصویر "خودروی آبی" توسط مدل تصویری) مثل ViT یا (ResNet به یک بردار ویژگی (v_i) در فضای جاسازی تبدیل می‌شود.
- توصیفات متنی نیز از طریق مدل متنی) مثل (GPT به بردارهای متنی (t_j) در همان فضای جاسازی نگاشت می‌شوند.

2. **محاسبه شباهت:**

مدل CLIP شباهت بین هر بردار تصویر و متن را با **ضرب داخلی کسینوسی** محاسبه می‌کند:

$$S(i,j) = (v\_i.t\_j) \ / \ (||v\_i|| \ ||t\_j||)$$

- شباهت بین تصویر و متن با توجه به همخوانی ویژگی‌های تصویری (مثل رنگ و نوع شیء) و ویژگی‌های متنی ارزیابی می‌شود.

ب) پیش‌بینی امتیازات شباهت:

1. **"یک خودروی آبی در جاده‌ای خلوت:"**
   - این توصیف احتمالاً بالاترین امتیاز شباهت را خواهد داشت.
   - دلیل: متن به ویژگی‌های اصلی تصویر، یعنی نوع شیء (خودرو) و رنگ آن (آبی)، اشاره دارد.
2. **"یک خودروی قرمز پارک‌شده در کنار ساختمان:"**
   - این توصیف امتیاز متوسطی دریافت می‌کند.
   - دلیل: نوع شیء (خودرو) با تصویر همخوانی دارد، اما رنگ "قرمز" با "آبی" متفاوت است و مکان نیز با تصویر تطبیق ندارد.
3. **"یک دوچرخه آبی در پارک:"**
   - این توصیف پایین‌ترین امتیاز را می‌گیرد.
   - دلیل: رنگ "آبی" با تصویر همخوانی دارد، اما شیء "دوچرخه" با "خودرو" تناقض دارد.

ج) تأثیر تغییر متن به "یک خودروی سبز در جاده‌ای شلوغ":

1. اگر متن به "یک خودروی سبز در جاده‌ای شلوغ" تغییر کند، امتیاز شباهت کاهش می‌یابد.
   - دلیل: ویژگی "سبز" با رنگ "آبی" در تصویر تطبیق ندارد.
2. **رفتار در فضای جاسازی:(Embedding Space)**
   - مدلCLIP ، تصاویر و متون مشابه را به بردارهایی نزدیک به یکدیگر در فضای جاسازی نگاشت می‌کند.
   - توصیف "خودروی سبز" و "خودروی آبی" به دلیل اشتراک در مفهوم "خودرو" ممکن است نسبتاً نزدیک باشند، اما تصویر "خودروی آبی" در فضای جاسازی به بردار توصیف "خودروی آبی" نزدیک‌تر است.

.5

الف) مقایسه مکانیزم Pooling Attention باGlobal Average Pooling

1. Global Average Pooling (GAP) :

- در این روش، تمامی مقادیر ویژگی در هر کانال از نقشه ویژگی با یکدیگر جمع شده و میانگین آن‌ها محاسبه می‌شود. این میانگین به عنوان نماینده آن کانال انتخاب می‌شود.
- فرمول محاسبه به صورت زیر است

$$GAP(x) = 1/(H \times W)\ sum\_i\ sum\_j\ x\_ij$$

که H و W به ترتیب ارتفاع و عرض نقشه ویژگی هستند.

ویژگی‌ها:

- این روش ساده و سریع است و در انتهای شبکه‌های CNN برای کاهش ابعاد استفاده می‌شود.
- **محدودیت** GAP **:**اطلاعات مکانی (Spatial) را به‌طور کامل حذف می‌کند، زیرا همه پیکسل‌ها با وزن یکسان در نظر گرفته می‌شوند.

3. Pooling Attention:

- این مکانیزم از یک ماتریس وزن یادگرفته‌شده برای مشخص کردن اهمیت هر موقعیت در نقشه ویژگی استفاده می‌کند.
- به جای میانگین‌گیری ساده، مقادیر نقشه ویژگی با وزن‌هایی که از طریق مکانیزم Attentionیاد گرفته شده‌اند، ترکیب می‌شوند. فرمول آن به صورت زیر است:

$$A = softmax(QK^T/\ sqrt(d\_k)), \qquad Z = AV$$

که A ماتریس توجه (Attention) است و اهمیت هر موقعیت را نشان می‌دهد.

ویژگی‌ها:

- Pooling Attentionبرخلاف GAP به بخش‌های مهم‌تر تصویر وزن بیشتری می‌دهد.
- این روش توانایی بیشتری در حفظ اطلاعات مکانی و مدل‌سازی روابط پیچیده در نقشه ویژگی دارد.

مقایسه کلی:

- GAP سریع و ساده است اما اطلاعات مکانی را حذف می‌کند.

- Pooling Attention انعطاف‌پذیرتر است و می‌تواند وزن بیشتری به نواحی مهم تصویر اختصاص دهد.

ب) تعداد درایه‌های صفر در ماتریس لیبل :N×N

تعریف ماتریس لیبل:

- در یادگیریContrastive ، ماتریس L_ij مشخص می‌کند که آیا دو نمونه i وj به یک کلاس تعلق دارند یا نه :

L_ij = 1 if i == 1 else 0

- **محاسبه تعداد درایه‌های صفر:**
- تعداد کل درایه‌ها در ماتریس N^2 :
- تعداد درایه‌های 1:
این تعداد برابر با تعداد عناصر قطر اصلی ماتریس است، یعنی N.
- تعداد درایه‌های :000تعداد درایه‌های 0 :

N^2 – N

ج) ضعف Zero-Shot مدل CLIP و دلایل آن

عملکرد ضعیف در وظایف خاص:

- مدل CLIP در وظایفی که نیاز به تشخیص تفاوت‌های دقیق یا درک روابط پیچیده بین اشیاء دارند، ضعیف‌تر عمل می‌کند. به‌طور خاص:

1. **طبقه‌بندی جزئی‌نگر :(Fine-grained Classification)** مانند شناسایی دقیق گونه‌های مشابه یا مدل‌های مختلف یک خودرو.
2. **وظایف وابسته به روابط فضایی :**مانند تشخیص دقیق موقعیت اشیاء نسبت به یکدیگر در تصویر.

دلایل ضعف در این وظایف:

1. **تمرکز بر ویژگی‌های کلی:(High-level Features)**
مدل CLIP برای تطبیق تصویر و متن بیشتر روی ویژگی‌های کلی مانند رنگ، شکل و موضوع تمرکز دارد. این رویکرد در تشخیص تفاوت‌های ظریف ضعیف عمل می‌کند.

2. **نبود داده‌های خاص:**
مدل CLIP بر اساس داده‌های عمومی و بدون برچسب دقیق آموزش دیده است. در نتیجه برای وظایف خاص یا پیچیده بهینه نشده است.

3. **ضعف در مدل‌سازی روابط فضایی:**
مدل CLIP به جای تمرکز بر ارتباطات مکانی دقیق بین اشیاء، تنها بر شباهت کلی بین تصویر و متن تکیه دارد. این امر باعث می‌شود که در وظایفی که نیاز به تحلیل دقیق روابط فضایی دارند، عملکرد ضعیفی داشته باشد.