نام و نام خانوادگی:

رضا قربانی پاجی

شماره دانشجویی:

403206565

تمرین پنجم درس یادگیری ژرف

سوال 1-بخش 1

آ)

1. **سطح جهانی (Global-Level)**: با استفاده از تابع هزینه L_{CLS}، مدل یاد می‌گیرد که درک کلی و مفهومی از کل تصویر داشته باشد مانند تشخیص اینکه در تصویر یک سگ وجود دارد. این سطح بالا، درک زمینه و محتوای اصلی را تضمین می‌کند.
2. **سطح محلی (Patch-Level)**: با استفاده از تابع هزینه L_{Patch}، مدل مجبور می‌شود که برای هر پچ (بخش کوچک) از تصویر نیز یک بازنمایی دقیق و معنادار تولید کند که با بازنمایی متناظر آن در مدل معلم یکسان باشد.

این نظارت مستقیم بر روی پچ‌ها باعث می‌شود جزئیات ظریف مانند بافت موی سگ یا لبه‌های دقیق یک برگ که در فرآیند یادگیری کلی نگر ممکن است نادیده گرفته یا میانگین گیری شوند، اهمیت پیدا کرده و به طور فعال در مدل حفظ شوند. در واقع مدل یاد می‌گیرد که همزمان هم کل جنگل و هم تک‌تک درختان را ببیند.

ب)

مدل‌های Segmentation سنتی، به خصوص آن‌هایی که مبتنی بر CNN هستند، معمولاً دو محدودیت اصلی دارند که این تکنیک به رفع آن‌ها کمک می‌کند:

1. **از دست رفتن اطلاعات مکانی دقیق:** مدل‌های سنتی برای افزایش میدان دید و کاهش محاسبات، مکرراً از لایه‌های Pooling یا Strided Convolutions استفاده می‌کنند. این کار باعث کاهش Resolution نقشه‌های ویژگی شده و منجر به از دست رفتن جزئیات دقیق و ایجاد مرزهای تار در قطعه‌بندی نهایی می‌شود. تکنیک patch-level distillation با وادار کردن مدل به حفظ اطلاعات دقیق در سطح هر پچ، این افت کیفیت را جبران می‌کند.
2. **تمرکز بر ویژگی‌های غالب و نادیده گرفتن جزئیات ظریف:** هدف اصلی در بسیاری از مدل‌ها، بهینه‌سازی یک تابع هزینه کلی برای کل تصویر است. این باعث می‌شود مدل بیشتر روی ویژگی‌های غالب و بزرگ که برای تشخیص کلی شیء مهم هستند تمرکز کند و جزئیات کوچک‌تر و بافت‌های ظریف را نادیده بگیرد. تکنیک L_{Patch} با ایجاد یک سیگنال یادگیری برای هر بخش کوچک از تصویر، این محدودیت را مستقیماً هدف قرار داده و تضمین می‌کند که جزئیات محلی نیز به همان اندازه اهمیت داده شوند.

سوال 1-بخش 2

آ)

مدل این تصمیم را بر اساس معیار Attention Value می گیرد که توسط خود معماری ویژن ترنسفورمر محاسبه می‌شود. این معیار نشان‌دهنده اهمیت معنایی پچ در درک کلی تصویر است.

پچ‌هایی که مقدار توجه بالایی دریافت می‌کنند، معمولاً مربوط به object های اصلی، یا بخش‌هایی هستند که دارای اطلاعات کلیدی و جزئیات مهم می‌باشند. پچ‌هایی با مقدار توجه کم، نمایانگر پس‌زمینه (Background) کم‌اهمیت، نواحی تکراری یا نویز در تصویر هستند.

بنابراین، مدل به صورت خودکار یاد می‌گیرد که بخش‌های حاوی اطلاعات غنی را از بخش‌های غیرضروری تشخیص داده و آن‌ها را در اولویت قرار دهد.

ب)

۱. افزایش دقت با کاهش نویز: با حذف کردن توکن‌های مربوط به پس‌زمینه و نواحی بی‌اهمیت، مدل می‌تواند تمام ظرفیت محاسباتی خود را بر روی تحلیل بخش‌های مهم تصویر متمرکز کند. این کار با بهبود نسبت سیگنال به نویز، به یک درک عمیق‌تر و دقیق‌تر از object اصلی منجر می‌شود.

۲. افزایش سرعت و کارایی: پردازش توکن‌ها، به خصوص در لایه‌های عمیق ترنسفورمر، بخش اصلی بار محاسباتی را تشکیل می‌دهد. با حذف بخش قابل توجهی از توکن‌های ورودی قبل از ورود به Student، حجم محاسبات به شدت کاهش یافته و در نتیجه سرعت آموزش و Inference مدل به طور چشمگیری افزایش می‌یابد.

سوال 1-بخش 3

آ)

توانایی مدل در reconstruction بخش‌های حذف‌شده، مستقیما به معنای یک ابزار افزایش وضوح نیست، بلکه یک روش آموزشی قدرتمند است که مدل را مجبور به یادگیری ذات و ساختار عمیق objectها می‌کند. این قابلیت به دو دلیل اصلی در دنیای واقعی بسیار ارزشمند است:

1. مقاومت Occlusion: در دنیای واقعی، objectها اغلب به صورت ناقص دیده می‌شوند مثلاً چهره‌ای که بخشی از آن با دست پوشانده شده یا خودرویی که پشت یک مانع قرار دارد. مدلی که یاد گرفته باشد بخش‌های نادیده را بازسازی کند، می‌تواند کل object را از روی یک بخش آن استنتاج کرده و تحلیل دقیقی ارائه دهد.
2. مقاومت در برابر نویز و کیفیت پایین: تصاویر در کاربردهای واقعی مانند دوربین‌های امنیتی، تصاویر ماهواره‌ای یا عکس‌های گرفته شده در نور کم اغلب دارای نویز، تاری یا Resolution پایین هستند. مدلی که توانایی بازسازی دارد، یک پیش فرض قوی از ظاهر objectها دارد و می‌تواند حتی با وجود اطلاعات ناقص ، آن‌ها را به درستی تشخیص داده و تحلیل کند. این باعث می‌شود مدل در شرایط غیرایده‌آل بسیار قابل اعتماد‌تر عمل کند.

ب)

اشیاء و صحنه‌هایی که دارای ساختار قابل پیش‌بینی و الگوهای مشخص هستند، بیشترین بهره را از این قابلیت می‌برند. زیرا مدل می‌تواند با دیدن بخشی از الگو، کل آن را بازسازی کند. نمونه‌های از آن عبارتند از:

موجودات زنده: به خصوص چهره و اندام انسان و حیوانات که دارای آناتومی مشخص و ساختار ثابتی هستند (مثلاً جایگاه چشم‌ها، بینی و دهان).

اشیاء ساخته دست بشر: وسایل نقلیه، ساختمان‌ها، مبلمان و متن نوشتاری، همگی از قوانین طراحی و ساختارهای منظمی پیروی می‌کنند که یادگیری و بازسازی آن‌ها را آسان‌تر می‌کند.

اشیاء با بافت‌های تکرارشونده: مانند پارچه، چوب یا نمای ساختمان که دارای بافت‌ها و الگوهای قابل پیش‌بینی هستند.

درمقابل، صحنه‌های بسیار شلوغ، بی‌نظم یا کاملاً انتزاعی (مانند یک تابلوی نقاشی آبستره) که فاقد ساختار مشخص هستند، بهره کمتری ازاین قابلیت خواهند برد.

سوال 1-بخش 4

آ)

مرحله اول: ایجاد کارایی و چالش توسط Attention-Based Token Removal:

ابتدا، تکنیک حذف توکن، نواحی کم‌اهمیت و پس‌زمینه تصویر را حذف می‌کند. این کار فوراً دو نتیجه دارد اولا مدل را سریع و کارآمد می کند چون داده کمتری برای پردازش وجود دارد. دوما، یک چالش برای مدل ایجاد می‌کند، زیرا حالا باید با یک ورودی ناقص کار کند.

مرحله دوم: کسب مقاومت و دقت توسط Pixel-Level Reconstruction و Patch-Level Self-Distillation:

کسب مقاومت با Reconstruction: مدل تلاش می کند تا پیکسل‌های حذف شده را بازسازی کند. این وظیفه آن را مجبور به یادگیری ساختار ذاتی و عمیق objectها می کند و در نتیجه، آن را در برابر ورودی‌های ناقص و دارای Occlusion مقاوم می‌سازد.

کسب دقت با Self-Distillation: همزمان، مدل featureهای استخراج‌شده از بخش‌های مهم و باقی‌مانده را با ویژگی‌های یک مدل teacher پایدار مقایسه می‌کند . این کار تضمین می‌کند که تحلیل این بخش‌های کلیدی، بسیار دقیق و سرشار از جزئیات باشد.

درمجموع، این سه تکنیک مدلی را می‌سازند که ابتدا با حذف اطلاعات اضافه سریع می‌شود، سپس با بازسازی در برابر نقص مقاوم می‌شود و درنهایت با خود-تقطیر در تحلیل بخش‌های مهم دقیق عمل می‌کند.

ب)

درصورت حذف Patch-Level Self-Distillation:

ازدست رفتن شدید جزئیات دقیق. مدل دیگر فشاری برای یادگیری ویژگی‌های باکیفیت در سطح local حس نمی‌کند و دقت آن در کارهای حساس مانند Segmentation دقیق به شدت افت می کند.

درصورت حذف Attention-Based Token Removal (و درنتیجه حذف Reconstruction):

کاهش شدید سرعت و مقاومت. مدل بسیار کندتر عمل خواهد کرد زیرا مجبوراست کل تصویر را پردازش کند. همچنین، چون دیگروظیفه بازسازی را انجام نمی‌دهد، مقاومت آن در برابر نویز و Occlusion به میزان قابل توجهی کاهش می‌یابد.

درصورت حذف Pixel-Level Reconstruction:

کاهش مقاومت و درک ساختاری. مدل همچنان سریع خواهد بود، اما دیگر مجبور به یادگیری ساختار عمیق objectها برای پر کردن جاهای خالی نیست. درنتیجه، توانایی آن برای مقابله با تصاویر ناقص و دارای Occlusion در دنیای واقعی، ضعیف خواهد شد.

سوال 2

الف)

دلیل اصلی نیاز روش‌هایی مانند SimCLR به negative pairs، جلوگیری از یک مشکل اساسی به نام **Model Collapse** یا رسیدن به راه حل بدیهی است.

توضیح مشکل Model Collapse:

اگر مدل فقط وظیفه داشته باشد که representation، positive pairs را به هم نزدیک کند، ساده‌ترین و کم‌هزینه‌ترین راه برای رسیدن به این هدف، این است که برای تمام تصاویر ورودی، یک خروجی یکسان و ثابت تولید کند. برای مثال، شبکه یاد می‌گیرد که بدون توجه به اینکه تصویر ورودی، عکس یک گربه است یا یک ماشین، همیشه یک بردار ثابت مثلاً بردار صفر را خروجی دهد.

در این حالت، مدل به هدف خود یعنی نزدیک کردن جفت‌های مثبت به طور کامل رسیده است، اما representation خروجی کاملاً بی‌ارزش است، زیرا هیچ اطلاعات معناداری درباره محتوای تصویر در خود ندارد و نمی‌تواند تصاویر مختلف را از یکدیگر تمایز دهد.

نقش نمونه‌های نامشابه (Negative Samples):

نمونه‌های نامشابه به عنوان یک نیروی دافعه عمل می‌کنند. آن‌ها مدل را مجبور می‌کنند که:

۱. نمایش positive pairها را به هم نزدیک کند.

۲. نمایش نمونه‌های نامشابه را از یکدیگر دور کند.

این دو هدف متضاد، مدل را وادار می‌کند تا ویژگی‌های کلیدی و تمایزدهنده تصاویر را یاد بگیرد. مدل برای اینکه بتواند همزمان نمایش گربه را به گربه دیگر نزدیک کرده و از نمایش ماشین دور کند، باید بفهمد چه ویژگی‌هایی گربه را تعریف می‌کنند.

بنابراین، نمونه‌های منفی در روش‌های یادگیری مقابله‌ای (Contrastive Learning) مانند SimCLR ضروری هستند تا از فروپاشی مدل جلوگیری کرده و شبکه را به سمت یادگیری نمایش‌هایی غنی، معنادار و قابل استفاده برای کارهای دیگر (مانند طبقه‌بندی) سوق دهند.

ب)

روش BYOL برای جلوگیری از مشکل Model Collapse، به جای استفاده از نمونه‌های منفی، همانطور که دکتر سلیمانی در کلاس درس گفتند از یک معماری نامتقارن استفاده می‌کند.

این معماری از دو شبکه تشکیل شده است:

1. Online Network: این شبکه اصلی است که به طور فعال با استفاده از گرادیان‌ها آموزش می‌بیند. وظیفه آن، پیش‌بینی کردن خروجی شبکه هدف است. این شبکه یک بخش اضافه به نام پیش‌بینی‌کننده predictor دارد.
2. Target Network: این شبکه، هدف یادگیری را برای شبکه آنلاین فراهم می‌کند. نکته کلیدی در همین شبکه نهفته است.

دلیل اصلی که BYOL دچارفروپاشی مدل نمی‌شود، نحوه به‌روزرسانی وزن‌های شبکه هدف است. وزن‌های این شبکه مستقیماً ازطریق backpropagation به‌روز نمی‌شوند؛ بلکه یک Exponential Moving Average ازوزن‌های شبکه آنلاین هستند.

به زبان ساده، شبکه هدف همیشه یک نسخه کمی قدیمی‌تر و پایدارتر از شبکه آنلاین است.
شبکه آنلاین مجبوراست ویژگی‌های معناداری را یاد بگیرد تا بتواند خروجی یک هدف پایدار را پیش‌بینی کند.
وجود predictor درشبکه آنلاین نیزاین عدم تقارن را تقویت می‌کند و باعث می‌شود رسیدن به یک راه حل بدیهی برای مدل دشوارتر شود.

بنابراین، BYOL با ایجاد این عدم تقارن—که درآن شبکه آنلاین سعی درپیش‌بینی یک شبکه هدفِ کندتر و پایدارتردارد نیازبه نیروی دافعه‌ی حاصل ازنمونه‌های منفی را ازبین می‌برد و با موفقیت ازفروپاشی مدل جلوگیری می‌کند.

ج)

این تفاوت درعملکرد، مستقیماً ازمعماری این دو روش نشأت می‌گیرد: SimCLR یک روش contrastive است، درحالی که BYOL یک روش predictive است.

Batch Size:

- **SimCLR:** الگوریتم SimCLR برای یادگیری به شدت به **negative samples** متکی است. این نمونه‌های منفی از دیگرتصاویرموجود درهمان batch می‌آید. هرچه batch size بزرگترباشد، تعداد نمونه‌های منفی بیشتر و متنوع‌تر است و مدل بهترياد می‌گیرد. با کاهش batch size، تعداد نمونه‌های منفی به شدت کم می‌شود و سیگنال یادگیری برای SimCLR بسیارضعیف و ناکارآمد می‌شود که افت شدید دقت را به همراه دارد.
- **BYOL:** درمقابل، BYOL اصلاً ازنمونه‌های منفی استفاده نمی‌کند. مکانیزم یادگیری آن داخلی و براساس پیش‌بینی خروجی شبکه هدف توسط شبکه آنلاین است. ازآنجایی که این فرآیند به دیگرتصاویرموجود درbatch وابسته نیست، عملکرد BYOL نسبت به کاهش batch size بسیارمقاوم‌تراست.

Tramformations:

- **SimCLR:** وظیفه SimCLR این است که ازمیان تمام نمونه‌های یک batch، positive pair خود را پیدا کند. وقتی transformations شدید باشند، مثلاًرنگ تصویرکاملاً حذف شود، دو نسخه ازیک تصویرممکن است بسیارمتفاوت به نظربرسند. این کارپیدا کردن جفت صحیح را برای SimCLR بسیاردشوارو مستعد خطا می‌کند.
- **وظیفه هوشمندانه BYOL:** وظیفه BYOL پیدا کردن جفت نیست، بلکه پیش‌بینی کردن است. شبکه آنلاین باید یاد بگیرد که representation شبکه هدف را پیش‌بینی کند. حتی اگر یک نسخه ازتصویر رنگی و نسخه دیگر بی‌رنگ باشد، مدل یاد می‌گیرد که فارغ ازاین تغییرات، محتوای اصلی تصویرچیست. این وظیفه پیش‌بینی، ذاتاً مدل را وادار به یادگیری

د)

دلیل این رویکرد را می‌توان به سه بخش تقسیم کرد:

۱. Global-to-Global

نزدیک کردن نمایش برش‌های global به یکدیگر، هدف اصلی SSL است. این کارمدل را وادارمی‌کند تا ویژگی‌های پایدارو اصلی تصویر را، فارغ از augmentations، یاد بگیرد. برای مثال، مدل می‌آموزد که مفهوم کلی سگ درهردو برش global وجود دارد.

۲. Local-to-Global

نزدیک کردن نمایش برش‌های local به برش‌های global، به مدل می‌آموزد که یک جزء، متعلق به یک کل است. به عبارت دیگر، مدل یاد می‌گیرد که یک تکه کوچک ازتصویرمثلاً فقط چشم سگ باید با مفهوم کلی سگ درارتباط باشد. این کاربه درک زمینه‌ای کمک می‌کند.

۳.Local-to-Local

این بخش، کلید اصلی این روش است. دورکردن نمایش برش‌های کوچک ازیکدیگر، مدل را مجبورمی‌کند تا ویژگی‌های خاص هربخش را یاد بگیرد.

اگرقراربود نمایش تمام برش‌ها به هم نزدیک شوند، مدل ممکن بود یک راه حل Naive پیدا کند. یادگیری ویژگی‌های ساده‌ترین بخش تصویرمثلاً یک تکه خزقهوه‌ای رنگ و تعمیم آن به کل تصویر. دراین صورت، اطلاعات مربوط به بخش‌های پیچیده‌تر و متمایزترمانند چشم، بینی یا گوش ازبین می‌رفت.

با وادارکردن مدل به ایجاد تمایزبین نمایش چشم و نمایش دم، مدل یاد می‌گیرد که این‌ها بخش‌های متفاوتی ازیک مفهوم کلی هستند. این امرمنجربه تولید feature mapهای بسیاردقیق‌تری می‌شود.