بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۴-۰۳

مدرس: مهدیه سلیمانی

تمرین پنجم | ددلاین تمرین : ۲۹ خرداد

- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. **بخش تئوری این تمرین تاخیر مجاز ندارد.**
- در هر کدام از سوالات، اگر از منابع خارجی استفاده کرده‌اید باید آن را ذکر کنید. در صورت همفکری با افراد دیگر هم باید نام ایشان را در سوال مورد نظر ذکر نمایید.
- پاسخ تمرین باید ماحصل دانسته‌های خود شما باشد. در صورت رعایت این موضوع، استفاده از ابزارهای هوش مصنوعی با ذکر نحوه و مصداق استفاده بلامانع است.
- پاسخ ارسالی واضح و خوانا باشد. در غیر این صورت ممکن است منجر به از دست دادن نمره شود.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- در صورتی که بخشی از سوال‌ها را جای دیگری آپلود کرده و لینک آن را قرار داده باشید، حتما باید تاریخ آپلود مشخص و قابل اتکا باشد.
- محل بارگذاری سوالات نظری و عملی در هر تمرین مجزا خواهد بود. به منظور بارگذاری بایستی تمارین تئوری در یک فایل pdf با نام HW5_[First-Name]_[Last-Name]_[Student-Id].pdf و تمارین عملی نیز در یک فایل مجزای زیپ با نام HW5_[First-Name]_[Last-Name]_[Student-Id].zip بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.
- طراحان: علی بناییان، کیارش جولایی، مهدی جمالی‌خواه، ایمان احمدی

بخش نظری (۱۰۰ نمره)

پرسش ۱. DetailCLIP (۵۰ نمره)

معماری DetailCLIP که در شکل ۱ نمایش داده شده، برای وظایفی که نیازمند قطعه‌بندی دقیق تصویر هستند، طراحی شده است. برخلاف مدل‌های سنتی که ممکن است جزئیات ریز را نادیده بگیرند، این مدل با استفاده از سه تکنیک، دقت را حفظ کرده و در عین حال از نظر محاسباتی کم‌هزینه باقی می‌ماند.

۱. (Patch-Level Self-Distillation)

در این روش، بخش‌های کوچکتر تصویر (دانش‌آموزان) از بخش‌های بزرگتر (معلمان) یاد می‌گیرند. این رویکرد به حفظ جزئیات کمک می‌کند که در غیر این صورت ممکن است از بین بروند. با تمرکز بر این تفاوت‌های ظریف، مدل از تقسیم‌بندی‌های اشتباهی که مدل‌های دیگر می‌کنند دور می‌کند.

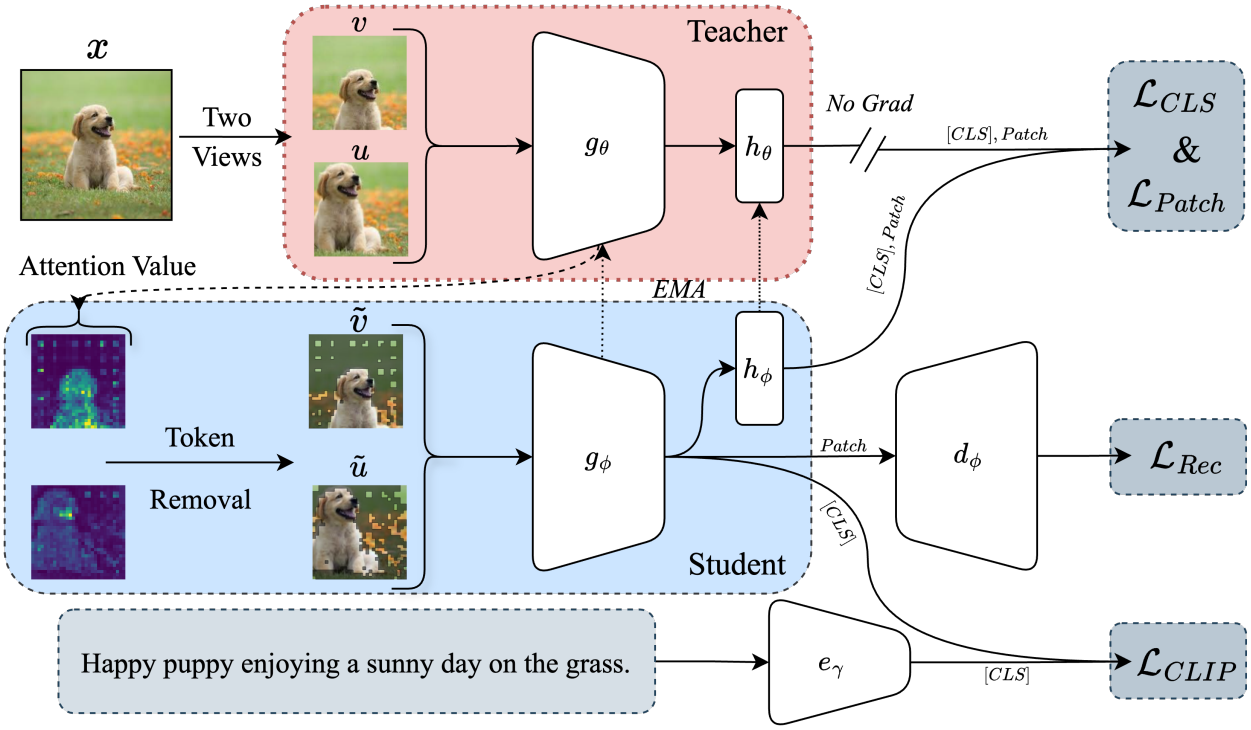
۲. حذف توکن (Attention-Based Token Removal)

این تکنیک مانند فیلتری برای داده‌ها عمل می‌کند. مدل تحلیل می‌کند که کدام بخش‌های تصویر اهمیت بیشتری دارد، برای مثال درک اهمیت شی داخل تصویر در برابر بکگراند تصور. این نه تنها سرعت تحلیل را افزایش می‌دهد بلکه با حذف نویز از مناطق غیرمهم، دقت را بهبود می‌بخشد.

۳. بازسازی در سطح پیکسل (Pixel-Level Reconstruction)

این روش برای افزایش وضوح تصویر به کار می‌رود. حتی هنگام کار با ورودی‌های با وضوح پایین، مدل می‌تواند خط‌های تیز و دقیقی را بازسازی کند. این ویژگی به‌ویژه برای اشکال پیچیده مانند خز حیوانات یا شاخ و برگ درختان ارزشمند است، جایی که لبه‌های دقیق برای تقسیم‌بندی دقیق ضروری هستند.

این روش‌ها با همدیگر همکاری می‌کنند تا مدل به دقت بالاتری برسد: Self-Distillation جزئیات ریز را حفظ می‌کند، حذف توکن پردازش را بهینه می‌کند، و بازسازی وضوح خروجی نهایی را دقیق‌تر می‌کند.



شکل ۱: با استفاده از معماری teacher-student، مدل teacher دو نمای مختلف از تصویر ورودی را پردازش کرده و مقادیر توجه (attention) را تولید می‌کند تا حذف توکن‌ها در مدل student را هدایت کند. سپس مدل student تصویر را با یک decoder vision بازسازی می‌کند، در حالی‌که هم‌زمان سه تابع هزینه شامل تابع طبقه‌بندی ($\mathcal{L}_{\text{CLS}}$)، تابع patch ($\mathcal{L}_{\text{Patch}}$)، و تابع بازسازی ($\mathcal{L}_{\text{Rec}}$) را بهینه می‌کند. همچنین تابع هزینه CLIP ($\mathcal{L}_{\text{CLIP}}$) به alignment میان انکودرهای تصویر و متن کمک می‌کند.

سؤالات:

۱. با توجه به پویایی teacher-student در یادگیری patch-level :

(آ) این رویکرد سلسله‌مراتبی چگونه به حفظ جزئیاتی از تصویر که ممکن است در غیر این صورت از بین بروند کمک می‌کند؟

(ب) مدل‌های دسته‌بندی سنتی چه محدودیت‌هایی دارند که این تکنیک به رفع آن‌ها کمک می‌کند؟

۲. درباره‌ی فیلترسازی (attention-based filtering):

(آ) مدل با چه معیارهایی تصمیم می‌گیرد کدام نواحی تصویر را در اولویت قرار دهد؟

(ب) این عمل (انتخاب اینکه به چه مکانی توجه شد) چه تأثیری بر کیفیت تحلیل دارد؟

۳. در مورد فرایند بازسازی (reconstruction):

(آ) چرا توانایی افزایش وضوح ورودی‌های کم‌کیفیت در کاربردهای دنیای واقعی ارزشمند است؟

(ب) کدام انواع اشیاء یا صحنه‌ها بیشتر از این قابلیت بهبود بهره‌مند می‌شوند؟

۴.

(آ) این سه مؤلفه چگونه یکدیگر را تکمیل کرده و در مجموع یک مدل قدرتمند می‌سازند؟

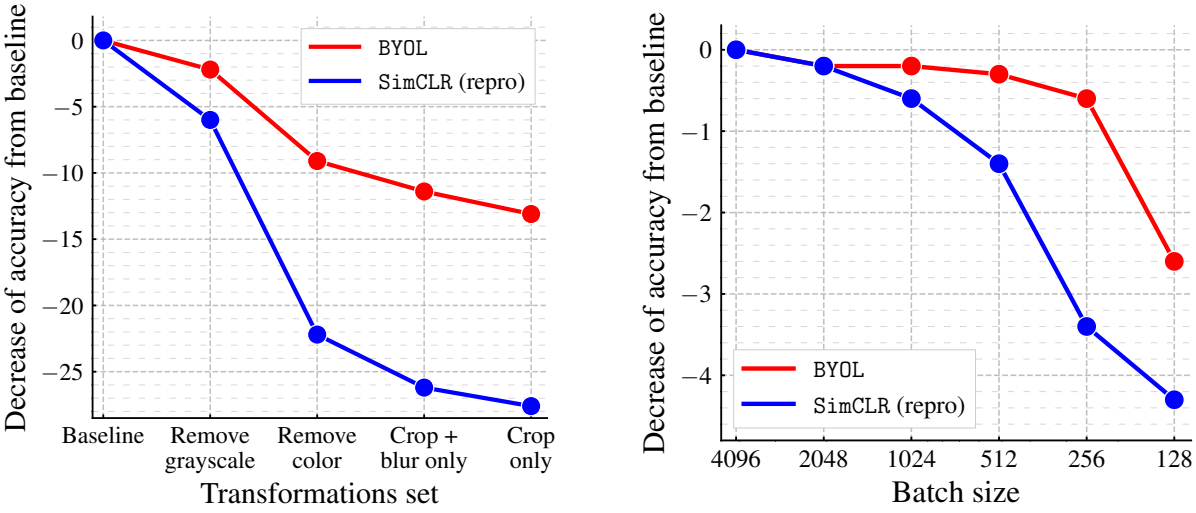
(ب) در صورت حذف یکی از این تکنیک‌ها، چه ضعف‌هایی ممکن است در عملکرد مدل ایجاد شود؟

پرسش ۲. یادگیری خودنظارتی (۵۰ نمره)

(الف) به طور کلی در روش‌های Self Supervised Learning تلاش بر این است که شبکه برای هر تصویر یک representation خروجی بدهد، به‌گونه‌ای که مفاهیم آن تصویر را در خود در بر داشته باشد. بسیاری از روش‌ها همچون روش‌هایی که در شکل می‌بینیم، این کار را با تلاش برای نزدیک‌کردن representation دو تصویر مشابه (positive pairs) انجام می‌دهند. با این حال چرا چرا برخی روش‌ها مانند SimCLR به نمونه‌ها نامشابه (negative samples) نیاز دارند تا خروجی مطلوبی داشته باشند؟

(ب) تحقیق کنید و بگویید دلیل اینکه چنین مشکلی در روش‌هایی چون BYOL رخ نمی‌دهد چیست.

(ج) در مقایسه‌ی BYOL می‌بینیم که این روش در برابر انتخاب برخی هایپرپارامترها همچون batch size و یا انتخاب transformationهایی که روی تصاویر اعمال می‌شوند مقاوم‌تر است (شکل ۲).



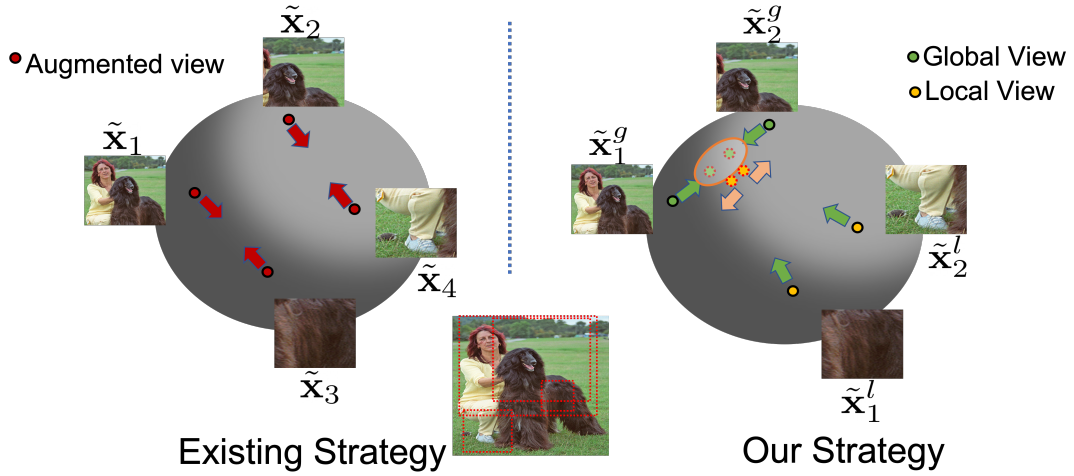
(ب) transformations removing of Impact

(آ) size batch of Impact

شکل ۲: BYOL: Bootstrap Your Own Latent

به نظر شما دلیل این موضوع چیست؟

(د) در برخی روش‌ها سعی می‌شود بازنمایی برش‌های بزرگ یک تصویر (global crops) به یکدیگر نزدیک و برش‌های کوچک local crops در عین حال که به بازنمایی برش‌های بزرگ نزدیک باشند، از یکدیگر دور شوند (شکل ۳).



شکل ۳: Representations. Global and Local Your Leverage

دلیل این موضوع را چه می‌دانید؟

پرسش ۳. (عملی) DINO

در این سوال می‌خواهیم با مدل DINO که به روش self-supervised ترین می‌شود، آشنا بشویم و در ادامه با gounded DINO که یک object detector بر پایه داینو هست کار کنیم.

پرسش ۴. (عملی) Stable Diffusion

هدف این سوال آشنایی با text-to-image generation با استفاده از مدل stable diffusion هست. ما گام به گام با نحوه ساخت عکس از روی یک عکس اولیه توسط این مدل آشنا می‌شویم و در ادامه با یک سری از مشکلات این مدل‌ها مانند object missing آشنا می‌شویم.

پرسش ۵. (عملی) Image Generation with CLIP

مدل‌های discriminative قابلیت یادگیری representation های دقیق از متن و تصویر به صورتی که یک شی امبدینگ‌های نزدیک در این دو modality داشته باشند را دارند. حالا در این تمرین می‌خواهیم از این قابلیت آن‌ها برای تولید عکس استفاده کنیم.