به نام خدا

یادگیری ژرف
نیم‌سال دوم ۰۳-۰۲
مدرس: دکتر مهدیه سلیمانی

سید امیر کسائی- ۴۰۲۲۱۲۲۱۴- همفکری با: امیر محمد عزتی

سوال اول:

۱.

آ)

- Negative samples نقش مهمی در بازنمایی در روش های contrastive learning مانند Momentum Contrast (MoCo) دارند. هدف از نمونه های منفی ارائه سیگنال متضاد به مدل در طول آموزش است. با ارائه مدل با جفت‌هایی از نمونه‌ها که در آن یکی مثبت در نظر گرفته می‌شود (به عنوان مثال، از یک تصویر یا زمینه یکسان سرچشمه می‌گیرد) و دیگری منفی در نظر گرفته می‌شود (مثلاً از یک تصویر یا زمینه متفاوت)، مدل یاد می‌گیرد که نمایش‌های جفت‌های مثبت را به هم نزدیک‌تر کند در حالی که نمایش‌های جفت‌های منفی را در فضای جاسازی دورتر از هم دور می‌کند. در اصل، نمونه‌های منفی به مدل کمک می‌کنند تا تمایز بین نمونه‌ها یا کلاس‌های مختلف را با به حداقل رساندن شباهت بین نمایش‌های جفت‌های منفی و در عین حال به حداکثر رساندن شباهت بین نمایش‌های جفت‌های مثبت، بیاموزد. این فرآیند به ایجاد نمایش‌های متمایزتر و معنادارتر کمک می‌کند که ساختار زیربنایی داده‌ها را به تصویر می‌کشد و منجر به بهبود عملکرد در مسئله های متفاوت مانند طبقه‌بندی تصویر، تشخیص اشیا و بازیابی تصویر می‌شود.
- در Bootstrap Your Own Latent (BYOL)، این روش بدون negative pair ها با معرفی یک چارچوب یادگیری منحصربفرد self-supervised که بر پیش‌بینی نسخه‌های قبلی خروجی‌های خود تمرکز دارد، کار می‌کند. این رویکرد شامل دو جزء کلیدی است: شبکه online و شبکه target. شبکه online برای پیش بینی نمایش شبکه target از همان تصویر تحت یک نمای augment شده‌ی متفاوت آموزش دیده شده است. به طور همزمان، شبکه target با slow-moving average پارامترهای شبکه آنلاین به روز می شود. این فرآیند تکراری پیش‌بینی نسخه‌های قبلی خروجی‌های خود و به‌روزرسانی شبکه هدف به تثبیت فرآیند یادگیری و بهبود representation بدون نیاز به جفت‌های منفی کمک می‌کند.

ب)

- در مدیریت اندازه های بزرگ batch در طول آموزش در SimCLR، رویکردهای زیر اتخاذ شد:
- **تغییر اندازه دسته ای**: اندازه batch آموزشی از ۲۵۶ به ۸۱۹۲ تنظیم شد، به این معنی که این اندازه در این مبازه تغییر میکند. با اندازه دسته بزرگتر ۸۱۹۲ تعداد قابل توجهی از نمونه های منفی در هر جفت مثبت از هر دو نمای تقویتی ارائه می شود.
- **انتخاب بهینه ساز:** برای اطمینان از ثبات در طول تمرین با اندازه های batch بزرگ، از بهینه ساز LARS برای همه اندازه های دسته استفاده کنند. این انتخاب برای مقابله با مسائل ناپایداری بالقوه ای که می تواند هنگام استفاده از SGD/Momentum استاندارد با مقیاس نرخ یادگیری خطی ایجاد شود، انجام شده است. این optimizer نقش مهمی در مدیریت فرآیند آموزش با تنظیم نرخ یادگیری برای پارامترهای مختلف، و تضمین بهینه‌سازی مدل پایدار و کارآمد بر اساس اندازه batch را دارد.

۲.

أ) در روشDINO ، شبکه teacher با استفاده از Exponential Moving Average (EMA) از شبکه student به‌روزرسانی می‌شود. این روش به شبکه teacher اجازه می‌دهد تا نمایشی پایدارتر و بدون نویز ایجاد کند. دلیل اصلی بهبود بازنمایی‌های شبکه teacher و بهبود انتقال دانش به شبکه student به شرح زیر است:

- پایداری و میانگین‌گیری: به‌روزرسانی شبکه teacher با استفاده از EMA باعث می‌شود که تغییرات شبکه student به‌طور نرم و تدریجی به شبکه teacher منتقل شود. این میانگین‌گیری نمایی تغییرات ناگهانی و نویزهای موقت را کاهش می‌دهد و بازنمایی‌های پایدارتری تولید می‌کند.
- تجمع دانش: با گذشت زمان، شبکه teacher با تجمع دانش از نسخه‌های مختلف شبکه student به بازنمایی‌های غنی‌تر و جامع‌تری دست می‌یابد. این تجمع دانش باعث بهبود بازنمایی‌ها و انتقال موثرتر دانش به شبکه student می‌شود.

ب) در فرآیند آموزشDINO ، شبکه teacher به سمت توجه به شی و نادیده گرفتن پس‌زمینه در تصویر تشویق می‌شود از طریق استفاده از تکنیک‌های augmentation قوی. این تکنیک‌ها شامل موارد زیر است:

- Augmentations قوی: تغییرات تصادفی در تصاویر ورودی مانند برش، تغییر رنگ، چرخش و اضافه کردن نویز باعث می‌شود که شبکه مجبور شود ویژگی‌های کلی و مهم اشیاء در تصویر را شناسایی کند و به جزئیات پس‌زمینه توجه نکند.
- Multi-Crop Strategy: در این روش، چندین نسخه برش خورده (crop) از یک تصویر ایجاد می‌شود و به مدل داده می‌شود. این نسخه‌ها شامل قسمت‌های مختلف تصویر هستند و شبکه باید یاد بگیرد که تمامی این نسخه‌ها مربوط به یک شیء مشترک هستند. این کار باعث می‌شود که شبکه به جای تمرکز بر جزئیات خاص، به ویژگی‌های کلی و اساسی شیء توجه کند.

ج)

- DINO:

۱. Contrastive Loss: استفاده از یک loss کنتراست (contrastive loss) که تضمین می‌کند بازنمایی‌های مشابه به هم نزدیک و بازنمایی‌های متفاوت از هم دور شوند. این کار باعث می‌شود که مدل بتواند تمایز بین بازنمایی‌ها را یاد بگیرد.
۲. EMA برایTeacher Network : به‌روزرسانی شبکه teacher با استفاده از EMA از شبکه student باعث می‌شود که بازنمایی‌ها به مرور زمان پایدارتر و غنی‌تر شوند.

- DINOv2:

۱. Enhanced Normalization and Regularization: استفاده از تکنیک‌های پیشرفته‌تر normalization و regularization برای بهبود استحکام و جلوگیری ازcollapse .
۲. Multi-Scale Features: بهره‌گیری از ویژگی‌های چندمقیاسی برای تضمین اینکه مدل بتواند بازنمایی‌های دقیق‌تری از تصاویر بدست آورد.

د) در مقاله DINOv2 ، نکات پیاده‌سازی شامل موارد زیر است:

- استفاده از Augmentations قوی و متنوع: برای تضمین یادگیری بازنمایی‌های متعادل و پایدار.
- Multi-Crop Strategy: استفاده از چندین برش از هر تصویر برای افزایش تنوع داده‌ها و بهبود یادگیری مدل.
- استفاده از Transformers: بهره‌گیری از معماری‌های Transformer به جای معماری‌های CNN برای بهبود قابلیت‌های مدل.
- Enhanced Normalization: استفاده از تکنیک‌های normalization برای بهبود استحکام و پایداری مدل
- Regularization Techniques: به کارگیری تکنیک‌های regularization برای جلوگیری از overfitting و بهبود تعمیم مدل.

ه)

- تفاوت‌های اصلی DINO با BYOL :

۱. Contrastive Loss در DINO :
   - DINO از یک loss کنتراست استفاده می‌کند که تضمین می‌کند بازنمایی‌های مشابه به هم نزدیک و بازنمایی‌های غیرمشابه از هم دور شوند. BYOL فقط از یک loss مبتنی بر مشابهت استفاده می‌کند که تنها بر نمونه‌های مثبت تمرکز دارد و نیازی به نمونه‌های منفی ندارد.

۲. Multi-Crop Strategy در DINO :
   - DINO از یک استراتژی multi-crop استفاده می‌کند که در آن چندین برش از یک تصویر به مدل داده می‌شود. این کار به مدل کمک می‌کند تا توجه بیشتری به شیء در تصویر داشته باشد و از پس‌زمینه چشم‌پوشی کند. در BYOL چنین استراتژی به کار نمی‌رود.

- تأثیر تفاوت‌ها در نتایج مدل DINO :
  - افزایش دقت بازنمایی‌ها : استفاده از contrastive loss در DINO باعث می‌شود که مدل بتواند بازنمایی‌های دقیق‌تر و با تمایز بیشتری یاد بگیرد.
  - تمرکز بیشتر بر ویژگی‌های شئ: استفاده از multi-crop strategy در DINO باعث می‌شود که مدل بتواند توجه بیشتری به ویژگی‌های کلی و مهم شیء در تصویر داشته باشد و از جزئیات پس‌زمینه چشم‌پوشی کند.
  - پایداری بیشتر: استفاده از EMA برای به‌روزرسانی شبکه teacher در DINO باعث پایداری بیشتر بازنمایی‌ها و تجمع بهتر دانش در طول زمان می‌شود.
  - این تفاوت‌ها به DINO کمک کرده‌اند تا نتایج چشمگیری در یادگیری بازنمایی‌های بدون نظارت به دست آورد.

سوال دوم:

۱.

أ) عمل جمع روی داده های همسایه ها و میانگین گیری روی آنها و استفاده از وزن یکسان برای همه همسایه ها ویژگی های permutation invariance و permutation equivariance را حفظ میکنند. بنابراین ای تابع معتبر است.

ب) عمل max به تنهایی مشکلی ایجاد نمیکند اما به دلیل استفاده از وزن های متفاوت برای همسایه ها، این تابع خواص permutation invariance و permutation equivariance را نقض میکنند پس معتبر نیست.

ج) با توجه به استفاده از max و همچنین وزن های یکسان برای همسایه ها، ها ویژگی های permutation invariance و permutation equivariance را حفظ میکنند. بنابراین ای تابع معتبر است.

۲.

$$t = 0, H_p^{(0)} = PH^{(0)}$$

$$GNN\ update\ rule: h_v^{(t+1)} = \sigma\left(Wh_v^{(t)} + \sum_{u\epsilon\mathcal{N}(v)} W'h_u^{(t)}\right)$$

در اینجا v'، نود های گراف اوریجینال است که به v در گراف جایگشتی map می‌شود. پس:

$$h_v^{(t+1)} = \sigma\left(Wh_{v'}^{(t)} + \sum_{u'\epsilon\mathcal{N}(v')} W'h_{u'}^{(t)}\right)$$

$$PH^{(t+1)} = H_p^{(t+1)}$$

$$PH^{(t+1)} = P\sigma\left(WH^{(t)} + AH^{(t)}W'\right)$$

$$Permuting\ the\ adjacency\ matrix\ A\ by\ P: A_p = PAP^T$$

$$H_p^{(t+1)} = \sigma\left(WH_p^{(t)} + A_pH_p^{(t)}W'\right)$$

$$Substituiting\ H_p^{(t)} = PH^{(t)}\ and\ A_p = PAP^T$$

$$H_p^{(t+1)} = \sigma(W\left(PH^{(t)} + (PAP^T)\left(PH^{(t)}\right)W'\right)$$

$$H_p^{(t+1)} = \sigma\left(PWH^{(t)} + PAP^TPH^{(t)}W'\right)$$

$$H_p^{(t+1)} = \sigma\left(PWH^{(t)} + PAH^{(t)}W'\right)$$

$$\sigma\ is\ element-wise: H_p^{(t+1)} = P\sigma\left(WH^{(t)} + AH^{(t)}W'\right)$$

$$H_p^{(t+1)} = PH^{(t+1)}$$

در واقع GNN برای هر جایگشتی از نود ها همگرا است که:

$$PH^{(t+1)} = H_p^{(t+1)}$$

سوال سوم:

۱. آشفتگی خصمانه فراگیر به آشفتگی‌های کوچک و نامحسوسی اطلاق می‌شود که وقتی به تصاویر طبیعی اضافه می‌شوند، می‌توانند باعث شوند طبقه‌بندی‌کننده‌های شبکه عصبی عمیق، بدون توجه به محتوای تصویر اصلی، تصاویر را با احتمال بالا طبقه‌بندی اشتباه کنند.

۲. کشف و مطالعه آشفتگی خصمانه فراگیر در شبکه‌های عصبی عمیق برای افزایش امنیت، درک تعمیم مدل، بهبود تفسیرپذیری و توسعه استراتژی‌های دفاعی در برابر حملات خصمانه مهم است. این آشفتگی‌ها آسیب‌پذیری‌ها را در طبقه‌بندی‌کننده‌ها آشکار می‌کنند، قابلیت‌های تعمیم را برجسته می‌کنند، بینش‌هایی را در مورد تصمیم‌گیری مدل ارائه می‌دهند، و پیشرفت‌هایی را در تحقیقات یادگیری ماشین خصمانه ایجاد می‌کنند آشفتگی خصمانه فراگیر نقش مهمی در کشف آسیب‌پذیری‌ها در طبقه‌بندی‌کننده‌های شبکه عصبی عمیق دارند و نشان می‌دهند که چگونه حتی تغییرات ظریف در داده‌های ورودی می‌تواند منجر به طبقه‌بندی اشتباه شود. با نشان دادن تعمیم‌پذیری این آشفتگی‌ها در تصاویر مختلف و معماری‌های شبکه عصبی، محققان بینش‌های ارزشمندی در مورد مرزهای تصمیم‌گیری اساسی طبقه‌بندی‌کننده‌ها به دست می‌آورند. این درک نه تنها به بهبود استحکام و تعمیم مدل کمک می کند، بلکه به توسعه سیستم های هوش مصنوعی قابل تفسیرتر کمک می کند. علاوه بر این، وجود آشفتگی فراگیر انگیزه اکتشاف استراتژی‌های دفاعی مؤثر در برابر حملات خصمانه را فراهم می‌کند و در نهایت زمینه یادگیری ماشینی متخاصم را پیش می‌برد و امنیت مدل‌های یادگیری عمیق را در کاربردهای عملی تقویت می‌کند.

۳. با توجه به مجموعه داده‌ای D متشکل از جفت‌های ورودی-خروجی (x_i, y_i)، که در آن x_i یک تصویر ورودی و y_i برچسب واقعی مربوطه است و یک تابع $g(x)$ که موفقیت حمله را اندازه‌گیری می‌کند، هدف ما یافتن یک آشفتگی فراگیر r به گونه ای است که وقتی به هر تصویر ورودی x اضافه می شود، باعث طبقه بندی اشتباه توسط شبکه عصبی شود.

فرض کنید $f(x)$ تابع طبقه بندی شبکه عصبی باشد و $x_i + r$ تصویر ورودی آشفته ای باشد که با اضافه کردن اغتشاش جهانی r به x به دست می آید. مسئله بهینه سازی را می توان به صورت زیر نوشت:

$$\max_r \sum_{(x_i,y_i)\in D} g(f(x_i + r))$$

به طوری که یک حد بالا برای نورم p آشفتگی r داشته باشیم که در آن $p \in [1, \infty)$:

$$||r||_p \leq \xi$$

در واقع میخواهیم احتمال خطای مدل در اثر آشفتگی را افزایش دهیم:

$$\mathbb{P}\big(f(x_i + r) \neq f(x_i)\big) \geq 1 - \delta$$

سوال چهارم:

۱. مقایسه مدل های SimVLM، CLIP و CoCa:

Feature	SimVLM	CLIP	CoCa
Goal	Understand text and image relationship	Understand text and image relationship	Understand text and image relationship
Integration of Vision & Language	Yes	Yes	Yes
Pre-training Approach	Yes	Yes	Yes
State of the Art Performance	Yes	Yes	Yes
Foundation Models	Yes	Yes	Yes
Architecture	Transformer-based	Dual Encoder	Transformer-based
Training Objective	PrefixLM	Contrastive Learning	Contrastive+Captioning
Zero-shot Learning	Limited capabilities	Strong zero-shot capabilities	Strong zero-shot capabilities
Performance	Strong performance on generative tasks	Strong performance on contrastive tasks	Strong performance on both generative and contrastive tasks

شباهت ها:

هر سه مدل ها، مدل های پایه‌ای هستند و برای فهمیدن رابطه بین متن و تصویر طراحی شده‌اند. هدف آنها بازنمایی مشترک عکس و متن برای انجام تسک های طبقه بندی تصویر، بازیابی تصویر و تولید تصویر است.

هر سه مدل هر دو حالت language و vision را در یک چارچوب یکپارچه ادغام می‌کنند و با پردازش مشترک متون و تصاویر تعاملات معنایی بین اطلاعات بصری و متنی را استخراج می‌کنند.

هر سه مدل ها از pre-training approach استفاده می‌کنند که ابتدا بر روی داده های زیادی آموزش داده می‌شوند تا بازنمایی های تعمیم یافته را یاد بگیرند. این موضوع برای ثبت روابط متنوع بین تصاویر و متون هائز اهمیت است.

هر سه مدل عملکرد پیشرفته‌ای در برخورد با مجموعه داده ها و تسک ها نشان داده‌اند و توانایی خود را در ثبت روابط پیچیده بین تصاویر و متون ثابت کرده‌اند. بنابراین این مدل ها برای برنامه های پیشرفته تر هوش مصنوعی کاندید های مناسبی هستند.

تفاوت ها:

مدل CLIP از contrastive learning استفاده می‌کند که شامل یادگیری بازنمایی با نزدیک تر کردن جفت های متن-تصویر مشابه و جدا کردن جفت های غیرمشابه است. این روش به پیدا کردن روابط معنایی بین تصویر و متن توسط مدل کمک می‌کند.

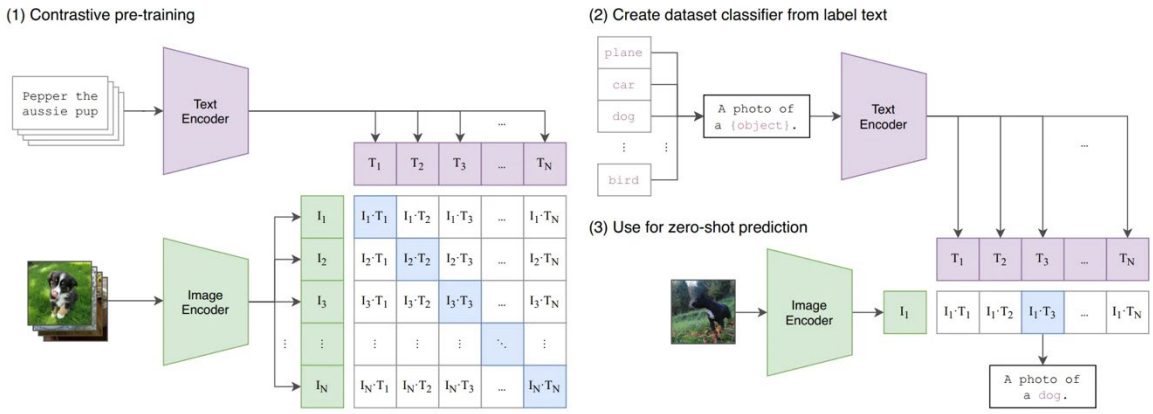
در حالی که SimVLM از PrefixLM استفاده می‌کند و CoCa ترکیبی از contrastive learning و generative task ها را استفاده می‌کند.

از لحاظ معماری، SimVLM و CoCa از معماری های مبتنی بر transformer ها استفاده می‌کنند اما CLIP از dual encoder ها استفاده می‌کند.

مدل های CLIP و CoCa برخلاف SimVLM در zero-shot learning خوب عمل می‌کنند و از لحاظ کارایی، SimVLM در تسک های generative، مدل CLIP در تسک های contrastive و مدل CoCa رو هر دو نوع تسک عملکرد خوبی دارند.

۲.

أ) بررسی مدل CLIP



- مدل CLIP از دو بخش اصلی تشکیل شده که شامل یک text encoder و یک image encoder است که به ترتیب وظیفه embed کردن متن و تصویر را دارند. برای text encoder معمولا از یک transformer با ساختار زیر استفاده می‌شود:
- ۶۳ میلیون پارامتر، ۱۲ لایه، ۵۱۲ لایه پنهان و ۹ تا attention head.این مدل بجای استفاده از وزن های آموزش دیده شده از ابتدا آموزش داده شده‌اند.
- برای image encoder، دو مدل مختلف ResNet-50 و Vision Transformer (ViT) را امتحان کردند. ResNet-50، از شبکه های عصبی کانولوشن استفاده می‌کند که برای image classification استفاده می‌شود. ViT اقتباسی جدید تر از transformer اصلی برای تصاویر است که در آن هر تصویر را می‌توان به دنباله‌ای از patch ها تقسیم کرد و به مدل داد. تحقیقات نشان داده که ViT سریع تر عمل train را انجام می‌دهد.

ب)

Loss function L_1:

$$L_1 = -\frac{1}{N}\log\left(\frac{e^{s_{ij}}}{\sum_{j=1}^{N} e^{s_{ij}}}\right) \text{ , where } s_{ij} = \frac{x_i.y_j}{\tau\|x_i\|\|y_j\|}$$

Simplifying L_1:

$$L_1 = -\frac{1}{N}\left(s_{ij} - \log\left(\sum_{j=1}^{N} e^{s_{ij}}\right)\right)$$

Gradient of $s_{ij} with\ Respect\ to\ x_i$:

$$N = x_i.y_j\ ,\ \ D = \tau\|x_i\|\|y_j\| \rightarrow \nabla_{x_i} s_{ij} = \frac{D.\nabla_{x_i}N - N.\nabla_{x_i}D}{D^2}$$

$$N = x_i.y_i \rightarrow \nabla_{x_i}N = y_j$$

$$D = \tau\|x_i\|\|y_j\| \rightarrow \nabla_{x_i} D = \tau\|y_j\|.\nabla_{x_i}\|x_i\|$$
$$\nabla_{x_i}\|x_i\| = \frac{x_i}{\|x_i\|} \rightarrow \nabla_{x_i} D = \tau\|y_j\|.\frac{x_i}{\|x_i\|}$$

$$\rightarrow \nabla_{x_i} s_{ij} = \frac{\tau\|x_i\|\left|\|y_j\|\right| y_j - (x_i.y_j)\tau\left|\|y_j\|\right|\frac{x_i}{\|x_i\|}}{(\tau\|x_i\|\|y_j\|)^2} = \frac{y_j}{\tau\|x_i\|\|y_j\|} - \frac{(x_i.y_j)x_i}{\tau\|x_i\|^3\|y_j\|}$$

Gradient of $L_1 with\ Respect\ to\ s_{ij}$:

$$\frac{\partial L_1}{\partial s_{ij}} = -\frac{1}{N}\left(1 - \frac{e^{s_{ij}}}{\sum_{j=1}^{N} e^{s_{ij}}}\right)$$

Chain rule:

$$\frac{\partial L_1}{\partial x_i} = \sum_{j=1}^{N} \frac{\partial L_1}{\partial s_{ij}} \frac{\partial s_{ij}}{\partial x_i}$$
$$\frac{\partial L_1}{\partial x_i} = \sum_{j=1}^{N}\left(-\frac{1}{N}\left(1 - \frac{e^{s_{ij}}}{\sum_{j=1}^{N} e^{s_{ij}}}\right)\right)\left(\frac{1}{\tau\|x_i\|\|y_j\|} y_j - \frac{x_i.y_j}{\tau\|x_i\|^3\|y_j\|} x_i\right)$$
$$\frac{\partial L_1}{\partial x_i} = -\frac{1}{N}\sum_{j=1}^{N}\left(1 - \frac{e^{s_{ij}}}{\sum_{j=1}^{N} e^{s_{ij}}}\right)\left(\frac{y_j}{\tau\|x_i\|\|y_j\|} - \frac{(x_i.y_j)x_i}{\tau\|x_i\|^3\|y_j\|}\right)$$