بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق
نیم‌سال دوم ۰۳-۰۲
مدرس: مهدیه سلیمانی

تمرین پنجم

- مهلت ارسال پاسخ تا ساعت **۲۳:۵۹** روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر تمرین‌های نظری بدون کسر نمره تا سقف ۵ روز و تمرین‌های عملی تا سقف ۱۰ روز وجود دارد. محل بارگزاری جواب تمرین‌های نظری بعد از ۳ روز و تمرین‌های عملی بعد از ۵ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد.
- هم‌فکری در انجام تمرین مانعی ندارد، فقط توجه داشته باشید که پاسخ تمرین حتما باید توسط خود شخص نوشته شده باشد. همچنین در صورت همفکری در هر تمرین، در ابتدای جواب تمرین نام افرادی که با آن‌ها هم‌فکری کرده اید را حتما ذکر کنید.
- برای پاسخ به سوالات نظری در صورتی که از برگه خود عکس تهیه می‌کنید، حتما توجه داشته باشید که تصویر کاملا واضح و خوانا باشد. درصورتی که خوانایی کافی را نداشته باشد، تصحیح نخواهد شد.
- 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 این تمرین : آقایان سامتی و هادیان

بخش نظری (۳۵ نمره)

سوال اول: (۱۵ نمره)

۱. با توجه به مقاله‌های SimCLR، MoCo و BYOL به سوالات زیر پاسخ دهید.

(آ) اهمیت وجود negative sampleها برای بدست آوردن بازنمایی چیست؟ چطور در روش BYOL این نیاز برطرف شده است؟

(ب) در روش SimCLR نویسندگان چه تمهیداتی را برای آموزش مدل با توجه به بزرگ بودن batch size در نظر گرفته‌اند؟

۲. با توجه به به مقاله‌های DINO و DINOv2 به سوالات زیر پاسخ دهید.

(آ) شبکه teacher دقیقا مشابه شبکه student و بدون آموزش پیشینی از ابتدا، به صورت Exponential Moving Average از شبکه student آپدیت می‌شود. در ابتدا چه چیزی باعث بهتر بودن بازنمایی بدست آمده از شبکه teacher و در نتیجه انتقال دانش به شبکه student می‌شود؟

(ب) در فرایند آموزش چگونه شبکه teacher به سمت توجه کردن به شئ و نادیده گرفتن پس زمینه در تصویر تشویق می‌شود؟

(ج) برای جلوگیری از collapse راهکار نویسندگان هر مقاله چیست؟

(د) نکات پیاده‌سازی در مقاله DINOv2 را مختصر توصیح دهید.

(ه) تفاوت‌های اصلی DINO با BYOL را مشخص کنید و توضیح دهید این تفاوت‌ها چطور در نتایج چشمگیر مدل DINO اثر داشته؟

پاسخ:

۱. (آ) برای جلوگیری از رسیدن شبکه به یک پاسخ بدیهی یعنی بدست آوردن بازنمایی ثابت بدون توجه به ورودی روش های SimCLR و MoCo .از نمونه‌های منفی برای تشکیل تابع هزینه استفاده می‌کنند در شبکه BYOL از یک شبکه target که وزن‌های آن به صورت Moving average از شبکه online آپدیت می‌شود و هم‌چنین وجود predictor در شبکه online از رسیدن به این پاسخ بدیهی جلوگیری می‌کند.

(ب) To keep it simple, we do not train the model with a memory bank. Instead, we vary the training batch size N from 256 to 8192. A batch size of 8192 gives us 16382 negative examples per positive pair from both augmentation views. Training with large batch size may be unstable when using standard SGD/Momentum with linear learning rate scaling. To stabilize the training, we use the LARS optimizer for all batch sizes. We train our model with Cloud TPUs, using 32 to 128 cores depending on the batch size.2 Global BN. Standard ResNets use batch normalization. In distributed training with data parallelism, the BN mean and variance are typically aggregated locally per device. In our contrastive learning, as positive pairs are computed in the same device, the model can exploit the local information leakage to improve prediction accuracy without improving representations. We address this issue by aggregating BN mean and variance over all devices during the training. Other approaches include shuffling data examples across devices, or replacing BN with layer norm .

۲. (آ) یک تحلیل در این رابطه با این سوال این است که با توجه به اینکه در شبکه teacher فقط از تصاویر global view استفاده می‌شود و شبکه student از local views هم استفاده می‌کند پس شبکه teacher اطلاعات بیشتری را در ورودی دریافت کرده و بازنمایی‌های غنی‌تری می‌تواند بدست آورد و به همین علت می‌تواند انتقال دانش به شبکه student داشته باشد.

(ب) معمولا در تصاویر global view به غیر از شئ، پس زمینه و اطلاعات دیگری نیز وجود دارد. اما در تصاویر local view معمولا تنها شئ و یا قسمتی از یک شئ در تصویر وجود دارد. با ورودی دادن این تصاویر به شبکه‌های teacher و student و در نهایت نزدیک کردن بازنمایی‌های بدست آمده شبکه teacher به توجه کردن به شئ در تصویر تشویق می‌شود.

(ج) در مقاله DINO علاوه بر استفاده از EMA در شبکه teacher برای جلوگیری از بازنمایی‌های یکسان بدون توجه ورودی، برای جلوگیری از یکسان شدن مقادیر در میان ابعاد بازنمایی و sharp،تر شدن در softmax یک عبارت temperature اضافه می‌کند. هم‌چنین برای غالب نشدن یکی از ابعاد در بازنمایی یک عبارت ثابت c با تمام ابعاد جمع می‌شود. جهت عملکرد مناسب بین بچ سایزهای مختلف این عبارت به صورت زیر محاسبه می‌شود.

$$c = mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_\theta(x_i)$$

در مقاله DIONv2 از Sinkhorn-Knopp (SK) batch normalization در شبکه teacher و از soft-max normalization در شبکه student استفاده می‌کند.

(د) بخش ۵ مقاله DINO

(ه) - Loss function در روش BYOL از MSE و در روش DINO از CrossEntorpy استفاده شده است. البته DINO با MSE Loss هم عملکرد به نسبت قابل قبولی دارد.
- predictor head در روش BYOL استفاده شده است و عملکرد روش در نبود این قسمت افت چشمگیری و برای جلوگیری از collapse کاربرد دارد دارد اما در DINO به این ماژول نیازی نیست و بودن یا نبودن آن تاثیر خاصی در عملکرد ندارد. در عوض از centering و sharpening در softmax استفاده شده است.
- multi-crop augmentation همانطور که در قسمت‌های قبل گفته شد بسیار در بدست آمدن بازنمایی‌ها در DINOو downstream taskها تأثیر گذار بوده اما در BYOL استفاده به تنهایی از multi-crop کارایی ندارد.

سوال دوم: (۸ نمره)

۱. برای یک گراف بدون جهت و بدون برچسب روی یال‌ها، تابعی که در هر لایه از یک شبکه عصبی گراف محاسبه می‌کنیم باید ویژگی‌های خاصی را رعایت کند تا بتوان از همان تابع (با اشتراک‌گذاری وزن‌ها) در گره‌های مختلف گراف استفاده کرد. فرض کنید برای یک گره مشخص i در گراف، $h_i^{\ell-1}$ پیام خودی (یعنی حالتی که در لایه قبلی برای این گره محاسبه شده است) برای این گره از لایه قبلی باشد، در حالی که پیام‌های لایه قبلی از n_i همسایه‌های گره i با $m_{i,j}^{\ell-1}$ نشان داده می‌شوند که j از ۱ تا n_i می‌باشد. ما از w با زیرنویس‌ها و بالانویس‌ها برای نشان دادن وزن‌های قابل یادگیری استفاده خواهیم کرد. اگر بالانویسی وجود نداشته باشد، وزن‌ها در سراسر لایه‌ها به اشتراک گذاشته می‌شوند. فرض کنید که همه ابعاد به درستی کار می‌کنند. توضیح دهید کدام یک از این‌ها توابع معتبر برای محاسبه پیام بعدی h_i^{ℓ} برای این گره هستند. برای هر انتخابی که معتبر نیست، به طور مختصر دلیل آن را ذکر کنید.

توجه: اعتبار به این معنی است که آن‌ها باید Invariance و Equivariance را که برای استفاده به عنوان یک GNN روی یک گراف بدون جهت نیاز داریم، رعایت کنند.

(آ) $h_i^{\ell} = w_1 h_i^{\ell-1} + w_2 \frac{1}{n_i} \sum_{j=1}^{n_i} m_{i,j}^{\ell-1}$

(ب) $h_i^{\ell} = \max(w_1^{\ell} h_i^{\ell-1}, w_2 m_{i,1}^{\ell-1}, w_3 m_{i,2}^{\ell-1}, \ldots, w_{n_i-1} m_{i,n_i}^{\ell-1})$ که در آن max به صورت مؤلفه‌ای بر روی بردارها عمل می‌کند.

(ج) $h_i^{\ell} = \max(w_1^{\ell} h_i^{\ell-1}, w_2 m_{i,1}^{\ell-1}, w_2 m_{i,2}^{\ell-1}, \ldots, w_2 m_{i,n_i}^{\ell-1})$ که در آن max به صورت مؤلفه‌ای بر روی بردارها عمل می‌کند.

۲. یک شبکه عصبی گراف (GNN) برای دسته‌بندی گره‌ها در یک گراف بی‌جهت $G = (V, E)$ در نظر بگیرید که در آن V مجموعه رئوس و E مجموعه یال‌ها است. GNN با استفاده از یک مکانیزم ارسال پیام به‌صورت تکراری ویژگی‌های گره‌ها را به‌روز می‌کند. فرض کنید $\mathbf{H}^{(t)}$ ماتریس ویژگی گره‌ها در تکرار t باشد که هر سطر $\mathbf{h}_v^{(t)}$ بردار ویژگی گره v را نشان می‌دهد.

عملیات ارسال پیام در تکرار t را به صورت زیر تعریف کنید:

$$\mathbf{h}_v^{(t+1)} = \sigma\left(\mathbf{W}\mathbf{h}_v^{(t)} + \sum_{u \in \mathcal{N}(v)} \mathbf{W}'\mathbf{h}_u^{(t)}\right),$$

که در آن:

(آ) $\mathbf{W}$ و $\mathbf{W}'$ ماتریس‌های وزن قابل یادگیری هستند،

(ب) $\mathcal{N}(v)$ مجموعه همسایگان گره v است،

(ج) σ یک تابع فعال‌سازی غیرخطی است.

ثابت کنید که GNN فوق نسبت به هر جایگشت گره‌ها هم‌وردا است. به عبارت دیگر، نشان دهید که برای هر ماتریس جایگشت $\mathbf{P}$، ویژگی‌های گره‌ها $\mathbf{H}^{(t)}$ پس از t تکرار، معادله زیر را ارضا می‌کند:

$$\mathbf{P}\mathbf{H}^{(t+1)} = \mathbf{H}_{\mathbf{P}}^{(t+1)}$$

که در آن $\mathbf{H}_{\mathbf{P}}^{(t+1)}$ ویژگی‌های گره‌هایی هستند که با اعمال همان عملیات GNN بر روی گراف جایگشت‌داده‌شده به دست می‌آیند.

پاسخ:

۱. (آ) This is valid because it is permutation invariant to the ordering of neighbors. This is the classic averaging form. Notice that a dependence on the number of neighbors is fine.

(ب) This is invalid. Since different scalar weights are applied to different m, it is not permutation invariant to the ordering of neighbors.

(ج) This is valid. Since the same weight w_2 is applied to all m, it is permutation invariant to the ordering of neighbors. The max is another classic permutation-invariant operation.

۲. Equivariance of Message-Passing: Applying the message-passing operation to $\mathbf{H}_{\mathbf{P}}^{(t)}$:

$$\mathbf{h}_v^{(t+1)} = \sigma\left(\mathbf{W}\mathbf{h}_v^{(t)}{}_{\mathbf{P}} + \sum_{u\in\mathcal{N}(v)} \mathbf{W}'\mathbf{h}_u^{(t)}{}_{\mathbf{P}}\right).$$

This can be written as:

$$\mathbf{H}_{\mathbf{P}}^{(t+1)} = \sigma\left(\mathbf{W}\mathbf{P}\mathbf{H}^{(t)} + \mathbf{P}\sum_{u\in\mathcal{N}(v)} \mathbf{W}'\mathbf{H}^{(t)}\right).$$

Since permutation is a linear operation and $\mathbf{P}$ commutes with matrix multiplication and summation:

$$\mathbf{P}\left(\mathbf{W}\mathbf{H}^{(t)} + \sum_{u\in\mathcal{N}(v)} \mathbf{W}'\mathbf{H}^{(t)}\right) = \mathbf{W}\mathbf{P}\mathbf{H}^{(t)} + \sum_{u\in\mathbf{P}(\mathcal{N}(v))} \mathbf{W}'\mathbf{P}\mathbf{H}^{(t)}.$$

Combining Results Applying the permutation after the message-passing operation gives the same result as applying the message-passing operation after the permutation:

$$\mathbf{P}\mathbf{H}^{(t+1)} = \mathbf{H}_{\mathbf{P}}^{(t+1)}.$$

Therefore, the GNN is equivariant to any permutation of the nodes.

سوال سوم: (۶ نمره)

با مطالعه‌ی مقاله‌ی Universal adversarial perturbations در مورد آشفتگی خصمانه‌ی فراگیر به سوالات زیر پاسخ دهید:

۱. به صورت خیلی مختصر و در حد یک الی دو جمله توضیح دهید که آشفتگی خصمانه‌ی فراگیر چیست.

۲. چرا یافتن چنین آشفتگی‌ای مهم است؟

۳. با داشتن داده‌های D و تابع g که میزان موفقیت حمله را اندازه‌گیری می‌کند (هر چه $g(x)$ بیشتر باشد، حمله موفق‌تر است) یافتن آشفتگی خصمانه‌ی فراگیر را به صورت یک مسئله‌ی بهینه‌سازی مقید به کرانی بر روی نرم p آشفتگی بنویسید.

پاسخ:

۱. آشفتگی خصمانه‌ی فراگیر، آشفتگی‌ای ثابت است که وقتی به تصاویر ورودی مدل اضافه می‌شود خروجی مدل اشتباه می‌شود. در واقع یک آشفتگی ثابت است که می‌تواند به تنهایی دقت مدل را به شدت کاهش دهد.

۲. یافتن چنین آشفتگی‌ای می‌تواند باعث شناخت ضعف‌های مدل مخصوصا در دنیای واقعی شود و بنابراین در جهت تقویت مدل در برابر حملات مفید واقع شود. حال این تقویت می‌تواند از جنس آموزش خصمانه یا موارد دیگر باشد. همچنین می‌توان از آن برای تفسیرپذیری نیز استفاده کرد.
دیگر کاربرد آن استفاده از آن برای ارزیابی مدل‌هاست.

۳.

$$\underset{v}{\text{argmax}}\ E_{x\sim D}[g(x+v)]\text{s.t. } \|v\|_p \leq \varepsilon$$

سوال چهارم: (۶ نمره)

۱. در درس با مدل‌های CLIP, SimVLM و CoCa آشنا شدید. دو شباهت و دو تفاوت این مدل‌ها را بیان کنید و موارد مربوطه را برای هر مدل به صورت واضح مشخص کنید.

۲. در مورد مدل CLIP به سوالات زیر پاسخ دهید:

(آ) ماژول‌های موجود در این مدل را نام برده و عملکرد هر کدام را توضیح دهید. برای هر کدام از این ماژول‌ها از چه مدلی استفاده شده است؟

(ب) می‌دانیم تابع هزینه‌ای که برای آموزش این مدل استفاده شده است از دو بخش تشکیل شده است که بخش اول آن به شکل زیر است:

$$\mathcal{L}_1 = -\frac{1}{N}\log\frac{e^{sim(x_i,y_i)/\tau}}{\sum_{j=1}^{N} e^{sim(x_i,y_j)/\tau}}$$

در صورتی که داشته باشیم $s_{ij} = \frac{x_i.y_j}{\tau\|x_i\|\|y_j\|}$ ، مشتق $\mathcal{L}_1$ نسبت به x_i یعنی $\frac{\partial\mathcal{L}_1}{\partial x_i}$ را محاسبه کنید.

پاسخ:

۱. شباهت‌ها:

(آ) هر سه مدل multimodal هستند و ورودی همزمان متن و تصویر را پشتیبانی می‌کنند..

(ب) هر سه مدل جزو مدل‌های بنیادی و از پیش آموزش دیده محسوب می‌شوند.

تفاوت‌ها:

(آ) معماری‌های متفاوت. مدل CLIP برای متن و تصویر از encoder های جداگانه استفاده می‌کند. مدل SimVLM از یک transformer واحد برای پردازش همزمان متن و تصویر استفاده می‌کند. مدل CoCa از VIT برای تصویر و transformer برای متن استفاده می‌کند.

(ب) نحوه آموزش. مدل CLIP از یک تابع هزینه‌ی contrastive استفاده می‌کند تا شباهت کسینوسی بین بازنمایی متن و تصویر را بیشینه کند. مدل SimVLM از پچ‌های تصویر به عنوان prefix token برای توکن‌های متن استفاده می‌کند و با استفاده از transformer و یک مدل encoder-decoder سعی در یادگیری همزمان متن و تصویر دارد. مدل CoCa به صورت ترکیبی از روش‌های دو مدل قبلی و همچنین MLM (Masked Language Modeling) استفاده می‌کند تا مزیت‌های هر دو را داشته باشد.

۲. (آ) Image Encoder: وظیفه‌ی تولید بازنمایی تصویر را به عهده دارد. معمولا از مدل‌های خانواده ResNet و یا ViT استفاده می‌شود.
Text Encoder: وظیفه‌ی تولید بازنمایی متن را دارد و از یک transformer ۱۲ لایه با causal mask استفاده می‌شود.

(ب)

$$\frac{\partial \mathcal{L}_1}{\partial x_i} = -\frac{1}{N}\frac{\partial}{\partial x_i}\left(\log \frac{e^{s_{ii}}}{\Sigma_{j=1}^{N} e^{s_{ij}}}\right) = -\frac{1}{N}\frac{\frac{\partial e^{s_{ii}}}{\partial x_i}.\Sigma_{j=1}^{N} e^{s_{ij}} \; - \; e^{s_{ii}}.\Sigma_{j=1}^{N}\frac{\partial e^{s_{ij}}}{\partial x_i}}{e^{s_{ii}}.\Sigma_{j=1}^{N} e^{s_{ij}}}$$

$$= -\frac{1}{N}\left(\frac{\frac{\partial e^{s_{ii}}}{\partial x_i}}{e^{s_{ii}}} - \frac{\Sigma_{j=1}^{N}\frac{\partial e^{s_{ij}}}{\partial x_i}}{\Sigma_{j=1}^{N} e^{s_{ij}}}\right) = -\frac{1}{N}\left(\frac{\partial s_{ii}}{\partial x_i} - \frac{\Sigma_{j=1}^{N}\frac{\partial s_{ij}}{\partial x_i}.e^{s_{ij}}}{\Sigma_{j=1}^{N} e^{s_{ij}}}\right)$$

$$\frac{\partial s_{ij}}{\partial x_i} = \frac{1}{\|x_i\|.\|y_j\|.\tau}\left(y_j - \frac{< x_i.y_j > x_i}{\|x_i\|^2}\right) = \frac{y_j}{\|x_i\|.\|y_j\|.\tau} \; - \; \frac{s_{ij}.x_i}{\|x_i\|^2}$$

$$\frac{\partial \mathcal{L}_1}{\partial x_i} = -\frac{1}{N\|x_i\|}\left(\frac{y_i}{\|y_i\|.\tau} - \frac{s_{ii}.x_i}{\|x_i\|} - \frac{\Sigma_{j=1}^{N} e^{s_{ij}}\left(\frac{y_j}{\|y_j\|.\tau} - \frac{s_{ij}.x_i}{\|x_i\|}\right)}{\Sigma_{j=1}^{N} e^{s_{ij}}}\right)$$

بخش عملی (۶۵ نمره)

توجه: لطفا در کلیه سوال‌های عملی نوت بوک تکمیل شده خود را به همراه سایر موارد در کوئرا بارگذاری کنید و از ارسال لینک و به اشتراک گذاری نوت بوک خودداری فرمایید.

سوال اول: (۲۰ نمره)

هدف از این سوال، آشنایی با مدل DINOv2 است. در این سوال از شما خواسته می‌شود با اضافه کردن یک لایه ترنسفورمر بر روی ویژگی‌های استخراج شده از داینو یک دسته‌بند بسازید. هدف این مدل دسته‌بندی تصاویر ماهواره‌ای برای تشخیص وجود یا عدم وژود پنل‌های خورشیدی است. همچنین در ادامه با استفاده از ماتریس توجه به دست آمده از آن می‌توانید اندازه پنل‌ها را تخمین بزنید.

سوال دوم: (۲۵ نمره)

نوت‌بوک StableDiffusion.ipynb شامل سه بخش پیاده‌سازی StableDiffusionPipeline به صورت Classifier-free guidance، اضافه کردن guidance اضافی برای تقویت رنگ آبی در تصویر تولیدی و در آخر fine tune کردن مدل برای آموزش یک مفهموم جدید شامل یک جفت توکن متنی و تصویر مربوطه به مدل با روش معرفی شده در مقاله DreamBooth است. در بخش آخر می‌توانید تصاویر مورد نظر خودتان را استفاده کنید.

توجه: نوتبوک برای محیط Google Colab بهینه شده است و توصیه می‌شود از این محیط برای تکمیل نوتبوک استفاده شود.

سوال سوم: (۲۰ نمره)

در این قسمت قرار است تا مباحث امنیت در یادگیری ماشین را به صورت عملی پیاده‌سازی کنید. برای انجام این بخش به نوت بوک `Adversarial_attacks_training.ipynb` مراجعه کنید. در این نوت بوک ابتدا دو نوع حمله FGSM و PGD را یکی با استفاده از کتابخانه و یکی به صورت from scratch پیاده‌سازی می‌کنید و سپس به کمک هردوی آن‌ها، آموزش خصمانه روی مدل انجام می‌دهید. تمامی بخش‌های داخل نوت‌بوک را تکمیل کنید و به سوالات مطرح شده پاسخ دهید.