دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش هوشمند تصاویر زیست پزشکی

نیم‌سال اول ۰۳-۰۲

مدرس: محمدحسین رهبان

پایان‌ترم | زمان آزمون: ۱۵۰ دقیقه

۱. (۱۰ نمره) تحلیل تصاویر میکروسکوپی

۱. (۴ نمره) پس از استخراج ویژگی‌ها توسط CellProfiler در یک سنجه میکروسکوپی و رسم نمودار tSNE، مشاهده کرده‌ایم که خوشه‌هایی تشکیل شده است که مرتبط با پروفایل داروهای متناظر با چاهک‌های نزدیک به هم از لحاظ مکانی است. به نظر شما چه مشکلی رخ داده و برای حل آن چه راه‌کاری پیشنهاد می‌کنید؟

۲. (۶ نمره) می‌خواهیم برای بررسی روی عملکرد ژن‌ها و تمایز، مشابهت و یا عملکرد ضد هم آن‌ها، از cell painting و morphological-profiling استفاده کنیم. برای این کار، می‌خواهیم در یک plate شامل تعدادی چاهک، عملکرد ژن‌های مختلف را بررسی کنیم. برای صحت‌سنجی این آزمایش چه کارهایی می‌توانیم انجام دهیم؟ (دو مورد ذکر کنید)

پاسخ:

۱. مشکل batch effect پیش آمده است؛ هر راه حلی که به صورت منطقی بتواند این اثر را از بین ببرد قابل قبول است. فقط توجه داشته باشید که همانطور که سر جلسه امتحان هم گفته شده است، اجازه تکرار آزمایش زیستی و یا طراحی و اجرای آزمایش زیستی دیگری را ندارید.

۲. neg-control: در بعضی از چاهک‌ها چیزی قرار نمی‌دهیم و این کار را برای حذف batch-effect انجام می‌دهیم. همچنین این کار کنترل می‌کند که بایاس سیستماتیکی در آزمایش اتفاق نیفتاده باشد.

pos-control: در بعضی چاهک‌ها آزمایشی را انجام می‌دهیم که خروجی آن و effect تولید شده را می‌دانیم و به این صورت کنترل می‌کنیم که آزمایش به درستی کار می‌کند یا خیر.

همچنین آزمایش را در چندین plate با شرایط کاملا یکسان انجام می‌دهیم برای اینکه مطمئن شویم که آزمایش و خروجی‌های که گرفته‌ایم reproducible هستند و به صورت اتفاقی به وجود نیامده‌اند.

۲. (۱۰ نمره) تفسیرپذیری

۱. (۴ نمره) در روش‌های attribution چرا گاهی ورودی مدل را در عددی کوچکتر از یک ضرب کرده و سپس گرادیان خروجی نسبت به ورودی را محاسبه می‌کنیم؟

۲. (۶ نمره) کیومرث یک روش attribution جدید طراحی کرد و درصدد ارسال روش پیشنهادی‌اش به کنفرانس ICLR است! کیومرث در ارزیابی روش خود، heatmap مربوط به تعداد از داده‌های آزمایش را مصورسازی کرده و نشان داده که این نقشه‌های گرمایی، روی نقطه مهم تصویر متمرکز شده است. شما به عنوان مشاور کیومرث چطور او را راهنمایی می‌کنید؟

پاسخ:

۱. در روش‌های attribution بعضی مواقع ممکن است x (ورودی) در ناحیه اشباع قرار گرفته باشد که به این ترتیب نمی‌توان گرادیان معنی داری برای آن به دست آورد. برای همین منظور با ضرب کردن این مقدار در یک عدد بین صفر و یک می‌توان اثر اشباع را تا جای ممکن حذف کرد و در نتیجه بررسی نمود که به نقاط معنی دار توجه شده است یا نه. (اسلایدهای تفسیرپذیری صفحه ۶۱ و ۶۲)

۲. بایستی مبتنی بر سه معیار مختلف مورد بررسی قرار بگیرد (اسلایدهای تفسیرپذیری صفحه ۷۵ تا ۷۸):

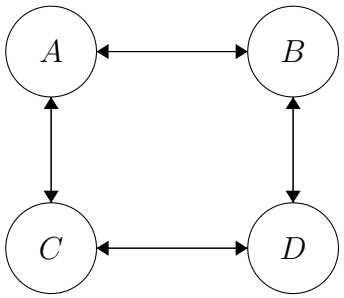
- معیار Coherence: چه میزان توانسته است بین ویژگی‌های متفاوت تصویر، تفاوت قائل شود.
- معیار Class Sensitivity: این معیار مشخص می‌کند که اگر در یک تصویر دو کلاس متفاوت باشند، وابسته به برچسب بیشتر به کدام قسمت توجه می‌کند.
- معیار Selectivity: بر اساس ارزش هربخش از تصویر در تصمیم‌گیری، سورت می‌شوند و سپس به ترتیب از پرارزش ترین به کم‌ارزش ترین حذف می‌کنیم تا ببینیم آیا اثر گذار در تغییر برچسب پیش‌بینی شده هست یا نه؟ بر این اساس می‌توان نمودار گفته شده در کلاس یعنی نمودار تعداد سمپل‌هایی که برچسب آن‌ها تغییر می‌کند نسبت به حذف پیکسل‌های ارزشمند را رسم کرد و نشان داد که روش ارائه شده چه مقدار نسبت به روش‌های دیگر مناسب است. مشابه همین کار را می‌توان با نقشه‌گرمایی نیز انجام داد و مشاهده کرد که چه قدر قسمت‌های انتخاب شده اثرگذاری بیشتری داشته‌اند.
به خصوص نیاز است نواحی‌ای که مدل به آن‌ها توجه داشته و در تصمیم‌گیری اثرگذار بوده‌اند بایستی گزارش شوند.

به صورت کلی انتظار می‌رود در این سوال ابتدا بحث شود که نواحی اثرگذار در تصویر بایستی توسط یک متخصص مشخص شده و مورد بررسی قرار گیرد و سپس نقشه‌گرمایی مورد تحلیل قرار گیرد که چه میزان توانسته است مدل مورد استفاده به این نقاط توجه کند. سپس معیارهای مختلف گفته شده در بالا بر روی روش ارائه شده مورد بررسی قرار گیرد و به خصوص معیار Selectivity بایستی توضیح داده شده و بررسی شود که روش پیشنهادی آیا واقعا یک روش تفسیرپذیری مناسب است یا خیر.

۳. (۳۰ نمره) مدل‌های دیفیوژنی و شبکه‌های عصبی مبتنی بر گراف

۱. با توجه به گراف زیر که در آن وضعیت‌های اولیه گره‌ها H^0 یک بعدی و به صورت زیر هستند به سوالات زیر پاسخ دهید:

$$H_A^0 = 1 \text{ ،} H_B^0 = 2 \text{ ،} H_C^0 = 3 \text{ ،} H_D^0 = 4$$



(آ) (۳ نمره) ابتدا ماتریس مجاورت گراف را بنویسید.

(ب) (۱۲ نمره) همچنین فرض کنید ماتریس وزن W یک ماتریس 1×1 با مقدار [2] و تابع فعال‌سازی σ تابع همانی یعنی همان $\sigma(x) = x$ باشد. وضعیت‌های به‌روز شده گره‌ها H^1 را محاسبه کنید.

۲. (آ) (۷ نمره) به چه دلیل می‌توانیم تابع هدف را که به صورت بیشینه کردن $\mathbb{E}_{q(\mathbf{x}_0)}[\log p_\theta(\mathbf{x}_0)]$ است را معادل کمینه سازی تابع هزینه

$$L_{\text{VLB}} = \mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big]$$

در نظر بگیریم؟

(ب) (۸ نمره) در رابطه با فرایند forward در مدل‌های دیفیوژنی، درستی رابطه زیر را بررسی کرده و توضیح دهید که $\bar{\alpha}_t$ در این رابطه چیست و چگونه محاسبه می‌شود؟

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{\bar{\alpha}_t} \mathbf{x}_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right)$$

پاسخ:

۱. ماتریس مجاورت A ، بردار وضعیت‌های اولیه H^0 و محاسبه H^1 به صورت زیر انجام شده است:

$$A=\begin{bmatrix}0&1&1&0\\1&0&0&1\\1&0&0&1\\0&1&1&0\end{bmatrix}\ H^0=\begin{bmatrix}1\\2\\3\\4\end{bmatrix}\ H^1=AH^0W=\begin{bmatrix}0&1&1&0\\1&0&0&1\\1&0&0&1\\0&1&1&0\end{bmatrix}\begin{bmatrix}1\\2\\3\\4\end{bmatrix}[2]=\begin{bmatrix}10\\10\\10\\10\end{bmatrix}$$

۲. پیش از هرچیزی لازم است اثبات کنیم تابع هدف ما (که مشابه مدل های VAE است) را میتوانیم به صورت کمینه سازی تابع هزینه معرفی شده بنویسیم. اینکار به صورت زیر انجام می شود:

$$\begin{aligned}-\log p_\theta(\mathbf{x}_0) &\leq -\log p_\theta(\mathbf{x}_0)+D_{\mathrm{KL}}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0))\\&=-\log p_\theta(\mathbf{x}_0)+\mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})/p_\theta(\mathbf{x}_0)}\Big]\\&=-\log p_\theta(\mathbf{x}_0)+\mathbb{E}_q\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}+\log p_\theta(\mathbf{x}_0)\Big]=\mathbb{E}_q\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big]\end{aligned}$$

۳. در این رابطه، $\bar{\alpha}_t$ نشان‌دهنده حاصل‌ضرب تجمعی $(1-\beta_t)$ تا گام tام است، جایی که β_t واریانس گام tام در فرآیند دیفیوژن است. به عبارت دیگر:

$$\bar{\alpha}_t=\prod_{s=1}^{t}(1-\beta_s)$$

برای اثبات این رابطه $\alpha_t=1-\beta_t$ و $\bar{\alpha}_t=\prod_{s=1}^{t}\alpha_s$ را تعریف می‌کنیم. حال با فرض اینکه

$$\epsilon_0,\epsilon_1,\ldots,\epsilon_{t-1}\sim\mathcal{N}(0,\mathbf{I})$$

می‌توانیم ترفند بازپارامتری‌سازی را به صورت بازگشتی برای نوشتن $\mathbf{x}_t$ به کار ببریم:

$$\begin{aligned}\mathbf{x}_t&=\sqrt{\alpha_t}\mathbf{x}_{t-1}+\sqrt{1-\alpha_t}\epsilon_{t-1}\\&=\sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2}+\sqrt{1-\alpha_{t-1}}\epsilon_{t-2})+\sqrt{1-\alpha_t}\epsilon_{t-1}\\&=\ldots\\&=\prod_{i=1}^{t}\sqrt{\alpha_i}x_0+\sqrt{1-\prod_{i=1}^{t}\alpha_i}\varepsilon_0\\&=\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\epsilon_0\end{aligned}$$

بنابراین، $\mathbf{x}_t$ از توزیع $\mathcal{N}(x_t;\sqrt{\bar{\alpha}_t}x_0,(1-\bar{\alpha}_t)I)$ پیروی می‌کند.

۴. (۵ نمره) مدل‌های مبدل

۱. (۳ نمره) برای تحلیل تصاویر پاتولوژی با ابعاد 224×224 از دو ویژن‌ترنسفورمر ViT-32 و ViT-16 استفاده شده است. نتایج ViT-32 بهتر از نتایج به دست آمده به کمک ViT-16 بوده است. با وجود این که ViT-16 دارای تعداد پچ بیشتری است، به نظر شما چرا چنین نتیجه‌ای حاصل شده است؟ چه trade-off ای وجود دارد که مانع عملکرد بهتر این مدل (ViT-32) شده است؟ (فرض کنید که نتایج به دست آمده برای هرکدام از آزمایش‌ها حاصل ۱۰۰ بار اجرا و میانگین گرفتن بین آن‌ها است. بنابراین نتایج کاملا دقیق است.)

۲. (۲ نمره) فقدان چه چیزی در مدل‌های ViT باعث می‌شود که به نسبت CNNها به تعداد داده بیشتر برای آموزش نیاز داشته باشند؟ توضیح دهید چگونه این نکته باعث می‌شود پس از یادگیری ViT نتایج بهتری از مدل‌های CNN داشته باشند؟

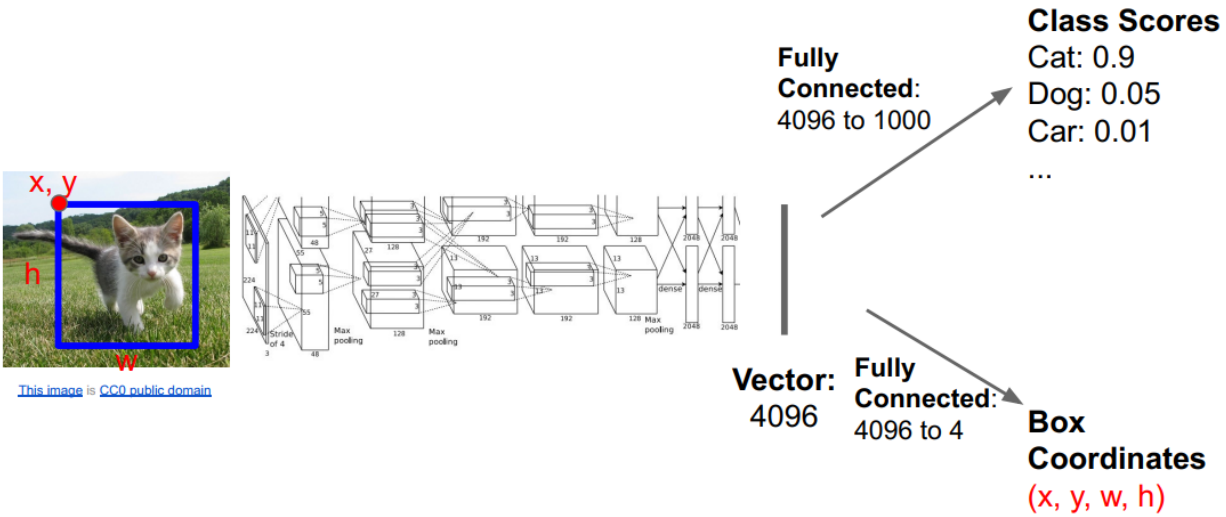
پاسخ:

۱.
- به دلیل receptive field بزرگ‌تر، ViT-32 می‌تواند در ثبت ویژگی‌های گسترده‌تر در یک تصویر مؤثرتر باشد.
- پچ‌های بزرگتر باعث کم شدن توجه به ویژگی های مهم محلی می‌شوند. چرا که هر کدام از پچ‌ها ابتدا به یک بردار امبدینگ تبدیل می‌شود. اگر ابعاد ورودی این تبدیل خطی خیلی زیاد باشد عملا خیلی اطلاعات مهمی درون یک پچ از بین می‌رود.

۲.
- در مدل‌های CNN روابط میان قسمت‌های مختلف از ابتدا به صورت محلی دیده می‌شود. چنین inductive bias ای می‌تواند به یادگیری سریع‌تر و بهتر مدل کمک کند. در نقطه مقابل، در مدل ViTروابط میان تمامی پچ‌های تصویر دیده می‌شود. (نه فقط با پچ‌های نزدیک) بدون داشتن inductive bias مدل‌های CNN این شبکه‌ها نیاز به داده یادگیری بیشتری دارند.
- همین امر از طرفی باعث می‌شود که ViTبتواند روابط بسیار غنی‌تری بین پچ‌هایی که حتی فاصله زیادی از نظر مکانی از هم دارند بیابد و نهایتا به دقت بالاتر منجر شود

۵. (۱۰ نمره) آشکارسازی اشیاء

۱. (۴ نمره)



با توجه به شکل بالا، برای آموزش هر شاخه از شبکه چه تابع هزینه ای باید داشته باشیم؟ (رابطه مربوط به تابع هزینه نوشته شود) همچنین توضیح دهید تابع هزینه‌ای که برای آموزش کل شبکه نوشته می‌شود باید به چه صورت باشد.

۲. (۶ نمره) معماری شبکه Region Proposal در شبکه Faster RCNN را شرح دهید (شامل معماری، داده‌های ورودی، داده‌های خروجی و تابع زیان).

پاسخ:

۱. برای آموزش شاخه‌ی بالایی که مربوط به دسته‌بندی است از لاس کراس آنتروپی و برای آموزش شاخه‌ی پایین از لاس MSE استفاده می‌شود. تابع هزینه‌ی کل شبکه برابر ترکیب خطی دو لاس ذکر شده در بالاست.

۲. در شبکه‌ی Faster R-CNN به جای استفاده از الگوریتم های سنتی مثل Selective Search از یک شبکه‌ی عصبی برای بدست آوردن Region Proposal ها استفاده می‌کنیم. به این منظور پس از اعمال یک شبکه‌ی کانولوشنی، بر روی فیچرمپ‌های بدست آمده به ازای هر نقطه، تعدادی باکس anchor با اندازه‌های مختلف می‌اندازیم . یک شبکه‌ی MLP وظیفه دارد برای هر باکس یک عدد به عنوان confidence وجود شی و ۴ عدد برای اصلاح Bounding Box هر anchor خروجی دهد. بنابراین داده‌های ورودی anchor ها و داده‌های خروجی به ازای هر anchor شامل ۵ عدد یعنی مشخصات باکس اصلاح شده و امتیاز وجود شی است. برای لاس این شبکه، از باینری کراس آنتروپی لاس برای امتیاز وجود شی و از MSE برای ۴ عدد مشخصه‌ی هر باکس استفاده می‌شود و لاس نهایی مجموع این دو لاس است.

۶. (۵ نمره) ثبت تصویر

۱. (۳ نمره) برای ثبت تصاویر CT-Scan که از اندام‌های شکمی گرفته شده است، از روش VoxelMorph استفاده کرده‌ایم. متاسفانه در زمان ارزیابی مشاهده کرده‌ایم که تصاویر دچار اعوجاج شده‌اند. چه تغییری در روش اعمال می‌کنید؟

۲. (۲ نمره) آیا برای آموزش شبکه VoxelMorph نیاز به دانستن نحوه انطباق واکسل‌های چند جفت تصویر به عنوان برچسب داده‌های آموزش داریم؟ شرح دهید.

پاسخ:

۱. تغییرات در روش‌های منظم‌سازی:

محدودیت‌های smoothness: استفاده از روش‌های منظم‌سازی L2 یا Total Variation که تغییرات مکانی زیاد در میدان تغییر شکل (deformation field) را جریمه می‌کنند. محدودیت‌های دیفئومورفیک: با جریمه کردن جزئیات با فرکانس بالا در میدان تغییر شکل، یا مثبت بودن دترمینان ماتریس ژاکوبی.

تغییرات در loss function:

ضرر مبتنی بر لندمارک: نشانگرهای آناتومیکی را برای هدایت فرآیند ثبت و جلوگیری از اعوجاج‌های غیرواقعی در نظر می‌گیرد. ثبت و سگمنتیشن مشترک: به طور مشترک با شبکه‌های سگمنتیشن ارگان آموزش می‌دهد تا از اطلاعات آناتومیکی برای alignmentبهتر استفاده کند.

تغییرات در معماری شبکه:

مکانیزم‌های توجه (Attention Mechanisms): از ماژول‌های Attention درون شبکه برای تمرکز بر روی نواحی مرتبط و کاهش اعوجاج‌ها در نواحی کم اهمیت استفاده می‌کند. معماری‌های U-Net: از معماری‌های مشابه U-Net برای حفظ اطلاعات فضایی (spatial) و جلوگیری از تغییر شکل‌های بیش از حد استفاده می‌کند.

۲. برای آموزش شبکه VoxelMorph، دانستن نحوه‌ی مطابقت دقیق واکسل‌ها در جفت‌های مختلف تصاویر به صورت صریح ضروری نیست. VoxelMorph به صورت یادگیری بدون نظارت تصاویر را ثبت می‌کند، به این معنا که نیازی به مطابقت‌های دستی انوتیشن شده بین تصاویر ندارد. به جای آن، شبکه یک تابع پارامتریزه شده (معمولاً یک CNN) را یاد می‌گیرد که با دریافت جفتی از تصاویر (یک تصویر ثابت و یک تصویر متحرک)، یک میدان تغییر شکلی را پیش‌بینی می‌کند که تصویر متحرک را به تصویر ثابت مطابقت می‌دهد.

۷. (۱۰ نمره) یادگیری چند نمونه‌ای

۱. (۶ نمره) مزایا و معایب دو رویکرد embedded-based و instance-based در یادگیری چند نمونه‌ای (MIL) چیست؟ برای حل این مشکل‌ها چه روشی پیشنهاد می‌شود؟

۲. (۴ نمره) چرا مشکل حافظه در روش‌های یادگیری چند نمونه‌ای به وجود می‌آید؟ راهکار حل این مشکل چیست؟

پاسخ:

۱. مزیت اصلی رویکرد embedded-based، جلوگیری از انتشار خطای شبکه تا مرحله تعیین برچسب برای هر پچ است. عیب اصلی آن نیز عدم تفسیرپذیری برای خروجی شبکه به دلیل مشخص نبودن برچسب هر پچ است. (۲ نمره)

برای رویکرد instance-based نیز برعکس رویکرد قبل، مزیت اصلی تفسیرپذیر بودن خروجی شبکه و عیب اصلی آن انتشار خطای موجود در شبکه تا مرحله تعیین برچسب است. (۲ نمره)

یک راه‌حل مناسب برای این مسئله، استفاده از روش Attention-based است. در این روش برای هر امبدینگ یک ضریب بدست می‌اوریم و با استفاده از این ضریب عملیات تجمیع نهایی برای ساختن امبدینگ تصویر اصلی را انجام می‌هیم. با این کار هم تفسیرپذیری حفظ می‌شود و هم انتشار خطا تا مرحله تعیین برچسب وجود نخواهد داشت. (۲ نمره)

۲. در روش MIL ما هر تصویر را که عموما تصویر بزرگی نیز به حساب می‌آید را به تعدادی پچ کوچک‌تر تقسیم می‌کنیم که به آن به اصطلاح bag می‌گوییم. زمانی که بخواهیم تمام این پچ‌ها را به عنوان ورودی به شبکه بدهیم و همچنین بخواهیم برای تمامی آن‌ها گرادیان حساب کرده تا خطای نهایی را محاسبه کنیم، به دلیل حجم زیاد پچ‌ها و بزرگ بودن شبکه با مشکل حافظه روبرو خواهیم شد. (۲ نمره)

یک راه‌حل مناسب که به ما نقطه شروع خوبی نیز برای آغاز آموزش مدل می‌دهد، استفاده از یک feature extractor از پیش آموزش دیده است. با این کار می‌توانیم این بخش از مدل را هنگام آموزش فریز کنیم و در نتیجه نیازی به نگهداری گرادیان برای آن نخواهد بود. در نتیجه حجم حافظه کمتری نیاز خواهیم داشت. (۲ نمره)

۸. (۵ نمره) تحلیل خطا

۱. (۳ نمره) در تحلیل تصاویر پاتولوژی برای یک بیمارستان خاص، متوجه شده‌ایم که تصاویر کمی حالت تارشدگی دارند. با این حال، ۵۰۰ تصویر برچسب‌گذاری شده در این بیمارستان جمع‌آوری کرده‌ایم. برای همین منظور خاص، ۵۰۰۰ تصویر در یک دادگان عمومی موجود در وب پیدا کرده‌ایم. نحوه تقسیم‌بندی این داده‌ها به آموزش، آزمایش و اعتبارسنجی به چه صورت باید باشد؟ چرا؟ (فرض کنید ۸۰ درصد کل داده‌ها برای آموزش، ۱۰ درصد برای آزمایش و ۱۰ درصد برای اعتبارسنجی تخصیص می‌یابد)

۲. (۲ نمره) دلیل نیاز به مجموعه training-dev چیست؟

پاسخ:

۱. توجه شود که ۵۰۰ داده‌ای که به بیمارستان تعلق دارد باید با اولویت بالاتر به مجموعه آزمایش و اعتبارسنجی تعلق گیرد. در کل ۵۵۰۰ داده داریم که ۸۰ درصد آن ۴۴۰۰ و ۱۰ درصد آن ۵۵۰ است. لذا تمام ۴۴۰۰ داده آموزش را از دادگان وب انتخاب می‌کنیم. برای ۵۵۰ داده آزمایش، ۲۵۰ داده از دادگان بیمارستان و ۳۰۰ داده از وب، و به شکل کاملا مشابه برای دادگان اعتبارسنجی عمل می‌کنیم.

۲. برای اینکه مطمئن شویم آیا روی دادگان آموزش به صورت ساده overfit شده‌ایم و عملکرد نامطلوب روی مجموعه اعتبارسنجی به دلیل تغییر توزیع نیست، از مجموعه training-dev استفاده می‌کنیم.

۹. (۱۵ نمره) یادگیری فدرال و محرمانگی تفاضلی

۱. (۵ نمره) فرض کنید ۱۲ بیمارستان با تجهیزات پردازشی کافی می‌خواهند بدون انتقال تصاویر پزشکی بین یکدیگر، مدلی با استفاده از یک روش یادگیری فدرال آموزش دهند. این مدل قرار است انواع یک بیماری را افتراق دهد. با توجه به اینکه هر کدام از این بیمارستان‌ها، مراجعه‌کننده بیشتری در یک نوع خاص از این بیماری را دارد، یک روش یادگیری فدرال با جزئیات کافی برای حل این مسئله پیشنهاد دهید.

۲. (۱۰ نمره) فرض کنید به عنوان مسئول پایگاه داده یک بیمارستان، می‌خواهید خدماتی را به یک مرکز پژوهشی ارائه کنید. جنس خدمات، محاسبه متوسط یک ستون دلخواه (که توسط کاربر مشخص می‌شود) به ازاء گروه‌های مختلف بیماران است. گروه‌های بیماران با یک ستون دیگر در پایگاه داده مشخص می‌شود. فرض کنید مقادیر ستون‌ها دارای محدوده مشخص هستند. به صورت خاص، ستون i ام حداقل برابر l_i و حداکثر برابر h_i است. تعداد بیماران در این پایگاه داده هم برابر ۱۰۰۰ است. اندازه هر گروه از بیماران هم حداقل ۵۰ نفر است. به منظور ایجاد محرمانگی تفاضلی با $\epsilon = \log 2$ از چه روشی استفاده می‌کنید؟ جزئیات روش را به صورت دقیق شرح دهید.

پاسخ:

۱. چون داده‌های بیمارستان‌ها iid نیستند باید از روش SCAFFOLD استفاده کنیم:

Algorithm Scaffold (server-side)

Parameters: client sampling rate ρ, global learning rate η_g

initialize θ, $c = c_1, \ldots, c_K = 0$
for each round $t = 0, 1, \ldots$ **do**
  $\mathcal{S}_t \leftarrow$ random set of $m = \lceil \rho K \rceil$ clients
  for each client $k \in \mathcal{S}_t$ in parallel **do**
    $(\Delta\theta_k, \Delta c_k) \leftarrow$ ClientUpdate(k, θ, c)
  $\theta \leftarrow \theta + \frac{\eta_g}{m} \sum_{k \in \mathcal{S}_t} \Delta\theta_k$
  $c \leftarrow c + \frac{1}{K} \sum_{k \in \mathcal{S}_t} \Delta c_k$

Algorithm ClientUpdate(k, θ, c)

Parameters: batch size B, # of local steps L, local learning rate η_l

Initialize $\theta_k \leftarrow \theta$
for each local step $1, \ldots, L$ **do**
  $\mathcal{B} \leftarrow$ mini-batch of B examples from $\mathcal{D}_k$
  $\theta_k \leftarrow \theta_k - \eta_l (\frac{\eta_k}{B} \sum_{d \in \mathcal{B}} \nabla f(\theta; d) - c_k + c)$
$c_k^+ \leftarrow c_k - c + \frac{1}{L\eta_l}(\theta - \theta_k)$
send $(\theta_k - \theta, c_k^+ - c_k)$ to server
$c_k \leftarrow c_k^+$

۲. قرار است تعدادی میانگین که از داده‌های متمایزی به دست آمده‌اند را اعلام کنیم پس نیازی به composition نداریم و هر میانگین را به صورت جداگانه بررسی میکنیم. برای ϵ-differential private کردن یک عدد حقیقی از مکانیسم لاپلاس استفاده می کنیم:

$$
\begin{aligned}
\bar{x}_{\text{clean}} &\sim \text{Laplace}(\bar{x}, \frac{\Delta f}{\epsilon}) \\
\Delta f &= \max_{\substack{\mathcal{D}_1, \mathcal{D}_2 \\ \text{neighbours}}} \|f(\mathcal{D}_1) - f(\mathcal{D}_2)\|_1 \\
&= \max_{x_n, \bar{x}^{(n-1)}, n} \|\frac{x_n - \bar{x}^{(n-1)}}{n}\|_1 \\
&= \frac{h - l}{50} \\
&\Rightarrow \bar{x}_{\text{clean}} \sim \text{Laplace}(\bar{x}, \frac{h - l}{50 \log 2})
\end{aligned}
$$