به نام خدا

یادگیری ژرف

نیم‌سال اول ۰۳-۱۴۰۲

مدرس: دکتر حمید بیگی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری پنجم (۱۶۰ نمره) | عنوان تمرین | زمان تحویل: ۸ دی (نظری)،۶ بهمن (عملی)

لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- 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 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 اعتنا باشد.
- تمام پاسخ‌های خود را در یک فایل با فرمت HW#_[SID]_[Fullname].zip روی کوئرا قرار دهید.

سوالات نظری (۹۰ نمره)

سوال ۱: بهینه سازی Generative Adversarial Networks (۲۵ نمره)

(آ) در این بخش هدف آشنایی با معیارهای فاصله دو توزیع است. روابط زیر را در نظر بگیرید:

- KL Divergence

$$D_{KL}(p\|q) = \sum_i p(i) \log\left(\frac{p(i)}{q(i)}\right)$$

- JSD Divergence

$$D_{JS}(p\|q) = \frac{1}{2} D_{KL}(p\|\frac{p+q}{2}) + \frac{1}{2} D_{KL}(q\|\frac{p+q}{2})$$

- Earth Mover's distance

$$W(p,q) = \inf_{\gamma \in \Gamma(p,q)} \mathbb{E}_{(x,y)\sim\gamma} \left[\|x-y\|\right]$$

$$\forall \gamma(x,y) \in \Gamma(p,q) : \int_y \gamma(x,y)\, dy = p(x), \quad \int_x \gamma(x,y)\, dx = q(y)$$

اکنون دو توزیع دو بعدی P ، Q را در نظر بگیرید.

$$\forall (x,y) \sim P, \quad x = 0, \quad y \sim \text{Uniform}(0,1)$$

$$\forall (x,y) \sim Q, \quad x = \theta\, (0 \le \theta \le 1), \quad y \sim \text{Uniform}(0,1)$$

۱. $JSD(P,Q), KL(Q||P), KL(P||Q)$ و $W(P,Q)$ را محاسبه کنید.

۲. دو توزیع P, Q دیگری بسازید که $JSD(P,Q)$ نسبت به پارامتر های P و Q مشتق‌پذیر نباشد.

(ب) یک شبکه GAN را در نظر بگیرید، در این شبکه بخش generator با دریافت یک بردار نویز $z \sim \mathcal{N}(0, I)$ سعی در تولید تصویر جعلی $G_\theta(z)$ دارد، همچنین بخش discriminator شبکه به صورت تصادفی با دریافت تصویر جعلی $G_\theta(z)$ یا یک تصویر واقعی از دادگان آموزشی، با استفاده از یک تابع sigmoid بر روی logit لایه آخر سعی در تشخیص واقعی یا جعلی بودن ورودی دارد:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$D_\phi(x) = \sigma(h_\phi(x))$$

برای آموزش شبکه‌های GAN دو تابع هدف $L_{\mathrm{D}}(\phi;\theta)$ و $L_{\mathrm{G}}(\theta;\phi)$ استفاده می‌شود. این توابع با استفاده از Gradient Descent به صورت دو مرحله‌ای بهینه‌سازی می‌شوند. یک مرحله $L_{\mathrm{D}}(\phi;\theta)$ در خلاف جهت گرادیان آپدیت می‌شود و در مرحله بعد $L_{\mathrm{G}}(\theta;\phi)$ آپدیت می‌شود.

$$L_{\mathrm{D}}(\phi;\theta) = \underbrace{-\mathbb{E}_{x\sim p_{\text{data}}}\left[\log D_\phi(x)\right]}_{\text{Real Data}} \underbrace{-\mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\log(1-D_\phi(G_\theta(z)))\right]}_{\text{Generated Data}}$$

$$L_{\mathrm{G}}(\theta;\phi) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\log(1-D_\phi(G_\theta(z)))\right]$$

۱. زمانی که در شبکه‌های GAN در بخش discriminator شبکه از backbone های pretrain استفاده شود این شبکه ها به سادگی تصاویر جعلی را شناسایی کرده و همین عملکرد خوب باعث vanishing gradient این شبکه‌ها خواهد شد. نشان دهید زمانی که خروجی discriminator برای تصاویر جعلی به صفر میل میکند:

$$D_\phi(G_\theta(z)) \approx 0$$

شبکه دچار vanishing gradient خواهد شد.(دقت کنید که در خروجی شبکه از تابع فعالساز sigmoid استفاده می‌کنیم. $D_\phi(x) = (\sigma(h_\phi(x))$

$$\hat{\nabla}_\theta L_{\mathrm{G}}(\theta;\phi) \to 0$$

۲. نشان دهید L_D هنگامی کمینه می‌شود که داشته باشیم $D_\phi = D^*$ که:

$$D^*(\phi) = \frac{p_{\text{data}}(x)}{p_\theta(x) + p_{\text{data}}(x)}$$

۳. برای یک generator ثابت θ و discriminator بهینه معادل D^*_ϕ نشان دهید که

$$V(G,D) = V(G,D^*_\phi) = \mathbb{E}_{x\sim p_{\text{data}}}\left[\log D_\phi(x)\right] + \mathbb{E}_{x\sim p_\theta}\left[\log(1-D_\phi(x))\right] = -\log 4 + 2\cdot \mathrm{D}_{\mathrm{JSD}}(p_{\text{data}}\|p_\theta)$$

(دقت کنید $D_{JSD}(p_{data}||p_\theta)$ معیار Jenson-Shannon Divergence است)

۴. ایده آل geneartor شبکه GAN تولید داده های جعلی، برگرفته از توزیع داده های آموزشی است($P_\theta \approx P_{data}$). همان طور که در بخش قبل دیدیم شبکه برای رسیدن به این هدف یک تابع هدف از جنس Jenson-Shannon Divergence استفاده می‌کند. و سعی دارد تا distance دو توزیع generator و توزیع داده های آموزشی را به حداقل برساند. معایب استفاده از JSD در مدل GAN را توضیح دهید.

۵. برای حل این مشکل مقاله Wasserstein GAN معیار Earth Mover's distance را معرفی میکند. با توجه به بخش های قبلی به صورت مختصر توضیح دهید که این معیار چرا از معیارهای KL Divergence و JS Divergence برای شبکه های GAN بهتر است.

سوال ۲: انرژی آزاد وردشی (۱۵ نمره)

روابط زیر را در نظر بگیرید.

$$\mathcal{F}(q) = \mathbb{E}_q[\log p(\mathbf{x}\mid\mathbf{z})] - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z}))$$

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z})) = \mathbb{E}_q[\log q(\mathbf{z}) - \log p(\mathbf{z})]$$

که $p(\mathbf{z})$ یک گاوسی چندمتغیره استاندارد است، یعنی:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I}) = \prod_{i=1}^{D} p_i(z_i) = \prod_{i=1}^{D}\mathcal{N}(z_i;0,1)$$

همچنین تقریب واریانسی $q(\mathbf{z})$ یک گاوسی چندمتغیره با ماتریکس کوواریانس قطری به فرم زیر است:

$$q(\mathbf{z}) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{D} q_i(z_i) = \prod_{i=1}^{D}\mathcal{N}(z_i;\mu_i,\sigma_i)$$

فرمول‌های زیر نیز برای یادآوری هستند:

$$\mathcal{N}(z;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$
$$\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right)$$

(آ) نشان دهید:

$$\mathcal{F}(q) = \log p(\mathbf{x}) - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z}\mid\mathbf{x}))$$

(ب) برای $\mathrm{D}_{\mathrm{KL}}\left(q_i\left(z_i\right)\|p_i\left(z_i\right)\right)$ یک فرمول صریح بر حسب μ_i^2 و σ_i^2 ارائه کنید.

(ج) نشان دهید:

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z})) = \sum_i \mathrm{D}_{\mathrm{KL}}\left(q_i\left(z_i\right)\|p_i\left(z_i\right)\right)$$

سوال ۳: مدل β-VAE (۱۵ نمره)

هدف از این سوال آشنایی با یک نمونه خاص از شبکه‌های VAE بنام β-VAE است. مقاله‌ی β-VAE را مطالعه کنید (در صورت آشنایی با VAE، خواندن این مقاله زمان کمی می‌گیرد).

(آ) تفاوت VAE و β-VAE را با توجه به ایده اصلی آن‌ها بیان کنید.

(ب) عبارت بهینه‌سازی این مدل در فرمول زیر آورده شده است. شرح دهید که مفهوم پشت این فرمول بهینه‌سازی چیست.

$$\max_{\phi,\theta}\mathbb{E}_{x\sim\mathbf{D}}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right]\right] \text{ به شرط } D_{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x})\|p(\mathbf{z})\right) < \epsilon)$$

(ج) تابع هزینه β-VAE در زیر آورده شده است. این تابع هزینه را از عبارت بهینه‌سازی قسمت قبلی سوال استنتاج کنید.

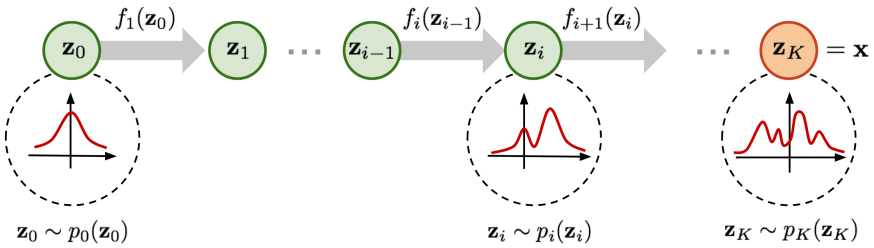
$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right] - \beta D_{KL}(q_\phi(\mathbf{z}\mid\mathbf{x})\|p(\mathbf{z}))$$

(د) طبق قسمت ۳ مقاله شرح دهید که علت اهمیت معیار تفکیک (Disentanglement Metric) چیست. همچنین شرح دهید که تغییر β چه تاثیری در آموزش شبکه می‌گذارد.

سوال ۴: جریان نرمال‌کننده (۲۰ نمره)

بخش اول

یک جریان نرمال کننده، یک توزیع ساده را با اعمال یک توابع تبدیل معکوس پذیر به یک توزیع پیچیده تبدیل می کند.



حال با توجه به نام گذاری های شکل بالا، می‌دانیم:

$$\mathbf{z}_{i-1} \sim p_{i-1}(\mathbf{z}_{i-1}) \quad (۱)$$
$$\mathbf{z}_i = f_i(\mathbf{z}_{i-1}) \quad (۲)$$
$$\mathbf{z}_{i-1} = f_i^{-1}(\mathbf{z}_i) \quad (۳)$$
$$p_i(\mathbf{z}_i) = p_{i-1}(f_i^{-1}(\mathbf{z}_i))|\det\frac{df_i^{-1}}{d\mathbf{z}_i}| \quad (۴)$$

با توجه به روابط بالا، برای اینکه جریان نرمال کننده باشد، باید تابع تبدیل از دو ویژگی زیر پیروی کند:

(آ) معکوس پذیر باشد.

(ب) دترمینان ژاکوبین آن محاسبه پذیر باشد.

حال، از قضیه قضیه تابع معکوس، زمانی که $y = f(x)$ باشد:

$$\frac{df^{-1}(y)}{dy} = \frac{dx}{dy} = (\frac{dy}{dx})^{-1} = (\frac{df(x)}{dx})^{-1} \tag{۵}$$

و همچنین می‌دانیم که دترمینان معکوس یک ماتریس معکوس پذیر معادل معکوس دترمینان آن ماتریس است زیرا:

$$\det(A)\det(A^{-1}) = \det(A \cdot A^{-1}) = \det(I) = 1 \tag{۶}$$

الف) حال در زمان *inference*، نیاز داریم که بتوانیم p_i را بر حسب z_{i-1} داشته باشیم، حال با توجه به دو گزاره بالا و عبارت $p_i(\mathbf{z}_i)$ از معادله ۴، رابطه $p_i(\mathbf{z}_i)$ را بر حسب z_{i-1} بازنویسی کنید و به رابطه زیر برسید:

$$\log p_i(\mathbf{z}_i) = \log p_{i-1}(\mathbf{z}_{i-1}) - \log \left| \det \frac{df_i}{d\mathbf{z}_{i-1}} \right|$$

ب) با استفاده از رابطه قسمت بالا، نشان دهید $\log p(x)$ برابر است با

$$\log p(\mathbf{x}) = \log \pi_0(\mathbf{z}_0) - \sum_{i=1}^{K} \log \left| \det \frac{df_i}{d\mathbf{z}_{i-1}} \right|$$

بخش دوم

حال، فرض کنید یک تبدیل خاص منظوره بصورت زیر داریم:

$$f : \mathbf{x} \mapsto \mathbf{y}$$

این تبدیل d بعد اول را تغییر نمی‌دهد و فقط از d+۱ تا D که سایز بردار ها هستند را دستخوش تبدیل می‌کند. بنابراین می‌توانیم آن را بصورت زیر فرمولیزه کنیم:

$$\begin{aligned} \mathbf{y}_{1:d} &= \mathbf{x}_{1:d} \\ \mathbf{y}_{d+1:D} &= \mathbf{x}_{d+1:D} \odot \exp(s(\mathbf{x}_{1:d})) + t(\mathbf{x}_{1:d}) \end{aligned}$$

که توابع s(.) و t(.) توابع تبدیلی هستند که هر دو چنین ویژگی ای دارند:

$$\mathbb{R}^d \mapsto \mathbb{R}^{D-d}$$

در دو بخش پایین می‌خواهیم بررسی کنیم که آیا می‌توان از این تابع در فرایند جریان نرمال کننده استفاده کنیم یا نه. بنابراین، باید دو ویژگی مورد نیاز را مورد بررسی قرار دهیم.
الف) نشان دهید که این تبدیل معکوس پذیر است.
ب) نشان دهید که دترمینان ژاکوبین آن محاسبه پذیر است و همچنین آن را بدست آورید.

سوال ۵: معیارهای ارزیابی مدل های مولد(۱۵ نمره)

برای ارزیابی مدل‌های مولد انتظار می‌رود معیارهای ارزیابی، مدل را از دو جهت ارزیابی کنند: اولا تصاویر تولید شده باید معنادار و با کیفیت باشند و ثانیا تصاویر تولید شده باید پراکندگی(diversity) داشته باشند.(برای نمونه معیار ارزیابی باید توانایی تشخیص مشکل mode collapse را داشته باشد.)

(آ) برای ارزیابی مدل های مولد تصویر از معیار های IS و FID استفاده می‌شود. این معیارها را به صورت مختصر توضیح دهید.

(ب) محبوب ترین معیار این حوزه FID است. با ارایه مثال معایب این معیار را بیان کنید.

(ج) یک از معیارهای دقیق‌تر که اخیرا معرفی شده معیار Density and Convergence است. با مطالعه مقاله به سوالات زیر پاسخ دهید.

۱. معیار percision و recall برای ارزیابی را مختصرا توضیح هید.

۲. این مقاله چگونه مشکل real outlier در percision و مشکل unrealistic divers samples در recall را حل می‌کند.

سوالات عملی (۷۰ نمره)

سوال ۶: Autoregressive Image Modeling (۲۵ نمره)

در این بخش یک مدل درست نمایی autoregressive را برای مدل سازی تصویر آموزش می دهیم. پیاده سازی ما بر مبنای PixelCNN خواهد بود و نقاط قوت و ضعف آن را بررسی خواهیم کرد. در انتهای نوتبوک بعد از آموزش مدل، ایراد اصلی معماری پیشنهادی که منجر به بهبود های پسین شده است را بررسی می کنیم و نهایتا سوالی در این زمینه پرسیده شده است که می بایست به آن پاسخ دهید.

سوال ۷: آموزش یک شبکه GAN برای تولید تصاویر ساده (۲۵ نمره)

در این تمرین، شما باید شبکه GAN و CGAN (Conditional GAN) را برای تولید تصاویر ساده ایجاد کنید. این تمرین به شما فرصت می‌دهد تا با مفاهیم اصلی GAN و نحوه پیاده‌سازی آن در PyTorch آشنا شوید. لطفاً مراحل اصلی پیاده‌سازی مولد و تمییزدهنده را در فایل HW5_Q7.ipynb انجام دهید. همچنین نتایج تولید تصاویر توسط مولد را نمایش دهید.

سوال ۸: شبکه خودرمزگذار متغیر کانولوشنی (۲۰ نمره)

در این سوال شما در ابتدا یک شبکه خودرمزگذار متغیر را پیاده‌سازی می‌کنید و سپس آن را با استفاده از داده آموزشی fashion-mnist آموزش داده و به تحلیل فضای نهان بدست آمده از آن می‌پردازید. با مراجعه به نوتبوک مورد نظر بخش‌های مشخص شده را تکمیل نمایید.