- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

سوالات نظری (۵ + ۶۰ نمره)

مسئله‌ی ۱. (۱۰ نمره)

(آ) یکی از راه‌های معمول برای تخمین گرادیان یک امید ریاضی، استفاده از رابطه‌ی زیر است:

$$\nabla_\theta \mathbb{E}_{z\sim q_\theta(z)}\left[f\left(z\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N} f\left(z^i\right).\nabla_\theta \ln q_\theta\left(z^i\right)$$

که در آن هر z^i نمونه‌ی مستقلی از توزیع $q_\theta\left(z\right)$ می باشد. درستی این رابطه را نشان دهید و بیان کنید که چطور می توانیم از آن در VAE استفاده کنیم. با مراجعه به این مقاله مشکلی که در استفاده از این روش وجود دارد را بیان کنید.

(ب) به صورت شهودی بیان کنید که روش Reparameterization چگونه می‌تواند این مشکل را حل کند؟

(ج) در بسیاری از موارد تابع خطای رمزگشای VAE را خطای MSE در نظر میگیریم. این در حالی است که هدف ما بیشینه کردن تابع $\mathbb{E}_{z\sim q_\theta(z|x)} \ln p_\theta\left(x|z\right)$ می باشد. در چه صورتی و با چه فرض هایی این دو کار معادل یکدیگر هستند؟

حل.

(آ) با نوشتن رابطه امیدریاضی و استفاده از قانون انتگرال لایبنیتز داریم:

$$\begin{aligned}\nabla_\theta \mathbb{E}_{zq_\theta(z)}\left[f\left(z\right)\right] &= \nabla_\theta \int_z f\left(z\right) q_\theta\left(z\right) dz \\ &= \int_z f\left(z\right) \nabla_\theta q_\theta\left(z\right) dz\end{aligned}$$

سپس با توجه به تساوی $\nabla_\theta q_\theta\left(z\right) = q_\theta\left(z\right) \nabla_\theta \ln q_\theta\left(z\right)$ داریم:

$$\nabla_\theta \mathbb{E}_{zq_\theta(z)}\left[f\left(z\right)\right] = \int_z f\left(z\right) q_\theta\left(z\right) \nabla_\theta \ln q_\theta\left(z\right) dz$$

سپس با تخمین مونته کارلو امیدریاضی بالا به مطلوب سوال دست پیدا می‌کنیم.
مشکلی که تخمین بالا دارد این است که با اینکه تخمینگر نااریبی می‌باشد ولی طبق نتایج تجربی واریانس بالایی دارد و تابع f در گرادیان تاثیر مستقیمی ندارد. برای کاهش واریانس ساده‌ترین روش افزایش مقدار N می‌باشد ولی باعث افزایش هزینه و محاسبات می‌شود.

(ب) باتوجه به اینکه هدف ما بیشینه‌سازی امیدریاضی تابع f می‌باشد، در تابع هزینه بالا گرادیان‌های بدست آمده براساس نمونه‌های بدست آمده از توزیع کدگذار می‌باشد و تابع f دخالت مستقیمی در بهینه‌سازی ندارد. به عبارت دیگر مسیر بهینه‌سازی و گرادیان های بدست آمده تنها وابستگی به نمونه‌های تولید شده دارد و ممکن است بعضی از نمونه‌هایی تاثیر مهمی دارند در فرایند نمونه‌گیری ظاهر نشوند. ولی در Reparameterization شرایط متفاوت می‌شود. شما فرض کنید توزیع کدگذار گاوسی می‌باشد. تابع هزینه بالا به صورت زیر تغییر پیدا می‌کند.

$$\nabla_\theta \mathbb{E}_{z\sim\, q_\theta(z)}\left[f\left(z\right)\right] = \nabla_\theta \mathbb{E}_{v\sim\, N(۰,۱)}\left[f\left(\sigma_\theta\left(x\right).v + \mu_\theta\left(x\right)\right)\right]$$

همانطور که در رابطه بالا مشخص است، در بدست آوردن گرادیان نسبت به پارامتر تتا، تابع f نیز تاثیر مستقیمی می‌گذارد و باعث می‌شود گرادیان های بدست آمده بهتر از حالت اول باشند.

(ج) در حالتی که فضای ما پیوسته باشد و توزیع داده شده در حالت کلی یک توزیع چندمتغیره گاوسی با واریانس I باشد، با جایگذاری در رابطه بالا یک عبارتی به شکل خطای MSE ظاهر می‌شود.

▷

مسئله‌ی ۲. (۱۵ نمره)

(آ) در یک VAE اگر داده ورودی از نوع باینری باشد (تصویری با پیکسل های ۰ و ۱)، می‌توان به جای توزیع گاوسی چندمتغیره روی خروجی کدگشا، از توزیع برنولی چندمتغیره استفاده کرد. توزیع خروجی کدگشا را به شکل برنولی چندمتغیره در نظر بگیرید و تابع فعال سازی آخرین لایه کدگشا را sigmoid در نظر بگیرید. اثبات کنید که در تابع هزینه این شبکه یک جمله ای به شکل Binary Cross Entropy ظاهر می شود.

(ب) تکنیک Reparameterization روی بسیاری از توزیع‌های پیوسته قابل اعمال است. تحقیق کنید که چگونه می‌توان از این تکنیک برای یک توزیع categorical استفاده کرد؟

(ج) یکی از نسخه‌های تغییریافته VAE مقاله مربوط به beta-VAE می باشد. در این روش یک ضریب beta ای پشت یکی از توابع هزینه قرار میگیرد. درباره این روش تحقیق کنید و بیان کنید:

۱) با انجام چه روندی از محاسبات، این ضریب در تابع هدف این روش ظاهر می‌شود؟

۲) هذف از افزودن ضریب beta چیست و اضافه کردن آن چه تاثیری روی ویژگی‌های فضای نهان یادگرفته‌شده توسط مدل دارد؟

حل.

(آ) فرض کنید ورودی $x \in \mathbb{R}^D$ به انکودر داده شده و z به دست آمده است. خروجی شبکه دیکودر را به صورت $a\left(z\right) = \sigma\left(f\left(z\right)\right) \in \mathbb{R}^D$ در نظر میگیریم و خروجی $\hat{x}$ را با نمونه برداری از توزیع $a\left(z\right)$ میسازیم. در این صورت احتمال آن که در خروجی شبکه دیکودر پس از نمونه برداری x به دست آمده باشد را میتوان با $P_\theta\left(\hat{x} = x|z\right)$ نمایش داد. حال اگر برای $\hat{x}$ یک توزیع برنولی چند متغیره در نظر بگیریم، میتوان این احتمال را به صورت زیر نوشت:

$$P_\theta\left(\hat{x} = x|z\right) = \prod_{i=۱}^{D} \left(a_i\right)^{x_i} \left(۱ - a_i\right)^{۱-x_i}$$

رابطه فوق چیزی نیست به جز احتمال مشترک چند متغیر برنولی که در آن a_i درایه i ام بردار خروجی $a(z)$ میباشد. با توجه به این توضیحات، میتوان لگاریتم این عبارت را به صورت زیر نوشت:

$$\log P_\theta(\hat{x}=x|z) = \log\left[\prod_{i=1}^{D}(a_i)^{x_i}(1-a_i)^{1-x_i}\right] = \sum_{i=1}^{D} x_i \log a_i + (1-x_i)\log(1-a_i) = -BCE(x, a(z))$$

(ب) راه حلی که برای این مشکل پیشنهاد میشود، روش Gumble-Softmax است که یک روش تقریبی میباشد. این روش در چند گام انجام میشود که باعث میشود اولا اپراتور احتمالاتی از مسیر اصلی گراف محاسبات کنار رفته و محاسبات به صورت deterministic انجام شود و ثانیا با کنار گذاشتن اپراتورهای گسسته از مسیر اصلی گراف، بتوان backprop را به درستی انجام داد. برای این منظور فرض کنید یک مسئله نمونه برداری k کلاسه داریم و میخواهیم از توزیع داده شده نمونه بگیریم. همچنین فرض کنید logits هایی که میخواهیم از آن ها نمونه بگیریم را با $\{a_i\}_{i=1}^{k}$ نشان دهیم. در این صورت این الگوریتم پیشنهاد میکند تا گامهای زیر طی شود:

- در گام نخست k متغیر $\{k_i\}_{i=1}^{k}$ را به صورت مستقل از توزیع $uniform(0,1)$ نمونه میگیریم.
- متغیرهای $\{g_i\}_{i=1}^{k}$ را به صورت $g_i = -\log(-\log(u_i))$ تشکیل میدهیم. در این صورت هر یک از متغیرهای g_i یک توزیع Gumbel استاندارد دارند.
- تا این جا شاخه احتمالاتی تولید متغیر را از گراف اصلی جدا کردیم حالا لازم است تا این متغیرها را با logit هایی که از قبل داشتیم ترکیب کنیم. برای این منظور جملات b_i را به صورت $b_i = a_i + g_i$ میسازیم. میتوان نشان داد که متغیر تصادفی $j = \arg\max_i b_i$ دقیقاً همان توزیعی را دارد که ما به دنبال آن بودیم. اما هنوز یک مشکل دیگر باقی مانده و آن استفاده از اپراتور گسسته argmax در وسط شبکه است.
- برای حل مشکل این اپراتور گسسته، از یک روش تقریبی با کمک لایه Softmax استفاده میکنند. لایه Softmax علاوه بر بردار توزیع احتمالات ورودی، یک پارامتر دیگر نیز به عنوان ورودی دریافت میکند که دما (Temperature) نام دارد. نحوه اعمال این پارامتر روی ورودی به این صورت است که ابتدا همه دادههای ورودی به این پارامتر دما (λ) تقسیم میشوند و سپس از Softmax عادی استفاده میشود. پارامتر دما، اثر ویژهای روی شکل بردار خروجی دارد؛ اگر $\lambda = 1$ باشد که همان سافتمکس معمولی خواهد بود اما هرچه λ به صفر نزدیکتر شود، بردار به دست آمده در خروجی به یک بردار one-hot نزدیکتر میشود. در واقع با کوچک شدن این پارامتر به اندازه کافی، تنها درایه با بیشترین مقدار در ورودی برابر یک شده و بقیه به صفر خیلی نزدیک میشوند. همچنین با زیاد شدن λ به سمت بینهایت، مستقل از توزیع ورودی، یک توزیع یکنواخت در خروجی خواهیم داشت. لذا مطلوب ترین حالت همان است که λ تا جای ممکن کوچک انتخاب شود تا تابع argmax به خوبی تقریب زده شود. البته باید توجه کرد که خیلی کوچک گرفتن پارامتر دما میتواند باعث زیاد شدن واریانس گرادیان بازگشتی از این لایه شود.

(ج) در βVAE دقیقاً مشابه VAE هدف یادگرفتن فضای نهانی است که بتواند دادها را خوب تولید کند با این تفاوت که در βVAE تاکید بیشتری روی disentangle بودن فضای نهان صورت میگیرد. به طور دقیقتر، βVAE نیز همانند VAE به دنبال بهینه کردن خطای بازسازی به صورت زیر است:

$$\max_{\theta,\varphi} \mathbb{E}_{x\sim P_{Data}}\left[\mathbb{E}_{z\sim q_\varphi(z|x)}\left[\log P_\theta(x|z)\right]\right]$$

علاوه بر عبارت فوق، لازم است تا یک شرط دیگر به منظور ساده کردن فضای نهان اضافه کنیم. به عبارت دیگر مسئله بهینه سازی βVAE یک مسئله constrained است که شرط آن روی فضای نهان به صورت زیر خواهد بود:

$$\max_{\theta,\varphi} \mathbb{E}_{x\sim P_{Data}}\left[\mathbb{E}_{z\sim q_\varphi(z|x)}\left[\log P_\theta(x|z)\right]\right]$$
$$KL(q_\varphi(z|x)\,||P(z)) < \delta \Rightarrow KL(q_\varphi(z|x)\,||P(z)) - \delta < 0$$

مسئله فوق یک مسئله بهینه سازی مشروط است که حل کردن آن الزاماً ساده نیست. برای از بین بردن شرط، از تکنیک لاگرانژ و ضریب β (به عنوان یک هایپرپارمتر) استفاده میشود و قسمت شرط را وارد عبارت بهینه سازی میکنند:

$$\mathbb{E}_{x\sim P_{Data}}\left[\mathbb{E}_{z\sim q_\varphi(z|x)}\left[\log P_\theta\left(x|z\right)\right]\right]-\beta\left(KL\left(q_\varphi\left(z|x\right)||P\left(z\right)\right)-\delta\right)$$

حال توجه کنید که عبارت $\beta\delta$ یک مقدار مثبت ثابت و مستقل از پارامترهاست لذا در مجموع میتوان تابع ضرر زیر را برای شبکه نوشت که به عنوان تابع ضرر βVAE شناخته می‌شود:

$$L\left(\theta,\varphi\right)=-\mathbb{E}_{x\sim P_{Data}}\left[\mathbb{E}_{z\sim q_\varphi(z|x)}\left[\log P_\theta\left(x|z\right)\right]\right]+\beta KL\left(q_\varphi\left(z|x\right)||P\left(z\right)\right)$$

▷

مسئله‌ی ۳. (۱۰ نمره)

(آ) همانطور که می‌دانیم، دو مدل GAN و VAE از مهم ترین و شناخته شده ترین مدل‌های generative در یادگیری عمیق می‌باشند. یکی از مهم‌ترین کاربردهای آن‌ها، تولید تصاویر و data augmentation می‌باشد. در این صورت، با فرض استفاده از مجموعه داده یکسان و فرآیند آموزش نسبتا کامل، آیا به طور کلی کیفیت تصاویر تولید شده توسط یکی از این مدل‌ها بر دیگری برتری دارد؟ لطفا پاسخ خود را با دلیل و در صورت نیاز اثبات ریاضی تشریح نمایید. می‌توانید از این مقاله استفاده نمایید.

(ب) تابع ReLU یکی از پرکاربردترین توابع فعالسازی مورد استفاده در شبکه های یادگیری عمیق می‌باشد، اما در برخی کاربری‌های خاص همانند بخش Generative در روش GAN می‌تواند منجر به ایجاد مشکل در فرآیند آموزش شود، به عبارت دیگر شبکه مولد ما آموزش نخواهد دید. در این مورد استثنا، توصیه می‌شود به جای ReLU ، از Leaky ReLU به عنوان تابع فعالسازی استفاده گردد. لطفا علت بروز مشکل به هنگام استفاده از ReLU را به طور کامل توضیح داده و تشریح کنید که به چه علت استفاده از Leaky ReLU می‌تواند مشکل را حل کند.

حل.

(آ) به طور کلی نمی توان هیچ مدلی را برتر دانست، و بسته به کاربرد هر کدام می توانند مفید واقع شوند. هر دو مدل نقاط ضعف و قوت خود می باشند. هر یک دارای معماری خاص خود بوده بسته به اینکه کدام مدل از زیر مجموعه GAN یا VAE مورد استفاده قرار گیرد، کیفیت خروجی تغییرات بسیاری خواهد کرد. به عنوان مثال، هنگامی که از DCGAN استفاده کنیم، سیستم یادگیری به صورت غیر نظارتی خواهد بود، در حالی که در VAE هر دو ساختار نیمه_نظارتی و نظارت شده استفاده شده است. همچنین، اندازه گیری خطای DCGAN مشکل می باشد چرا که از مکانیزم minimax برای یادگیری استفاده کرده و خطای MSE آن در هر Epoch تغییرات محسوسی را دارد. در نتیجه، بسته به کاربرد و تعداد داده های موجود برای آموزش، عملکرد هر یک از مدل ها می تواند به دیگری برتری داشته باشد.

(ب) تابع relu همانطور که اشاره شد، یکی از پرکاربردترین توابع در زمینه توابع فعالسازی می باشد. اما مشکلی که در هنگام استفاده از این تابع در برخی موارد در شبکه های حساسی همانند generator می تواند ایجاد شود، عدم آموزش دیدن شبکه از جایی به بعد می باشد. علت این امر آن است که در تابع relu ، هنگامی که ورودی کوچکتر یا مساوی صفر باشد، خروجی صفر خواهد شد، و این بدان معناست که در شبکه هایی همانند generator در شبکه های GAN که آموزش به شدت حساسی دارند و گاهی آموزش شبکه بسیار کند پیش می رود، یعنی rate آموزش حوالی صفر است، با استفاده از relu عملا آموزش متوقف شده و شبکه خروجی مطلوب را تولید نخواهد کرد. از آن جایی که توابعی همانند relu leaky در حوالی صفر و مقادیر کوچکتر از آن، مقداری غیر از صفر دارند لذا شبکه با استفاده از آن ها حتی در زمان هایی که rate آموزش پایین است نیز شانس ادامه آموزش دیدن را دارد.

▷

مسئله‌ی ۴. (۱۵ نمره)

(آ) یکی از مشکلات شایع در شبکه های GAN مشکل Mode Collapse می باشد که باعث می شود شبکه GAN به یک حالت عدم آموزش رسیده و به طور متوالی خروجی های یکسان تولید کند. این مشکل را به طور کامل تشریح کرده و راه حل های احتمالی به منجر به غلبه بر این مشکل می شوند را بیان نمایید.

(ب) معماری کلی و تابع هزینه مدل W-GAN را تشریح نمایید و تفاوت های آن با مدل پایه GAN را توضیح دهید. آیا تابع هزینه این مدل کمکی به برطرف شدن مشکل Mode Collapse خواهد کرد؟ توضیح دهید.

حل.

(آ) پدیده Mode Collapse در شبکه های GAN ، زمانی اتفاق می افتد که قسمت generator تنها قادر به تولید یک نوع خروجی یا طیف بسیار محدودی از خروجی ها می باشد. این امر می تواند به دلیل آموزش نامناسب شبکه اتفاق بیفتد یا در حالتی که شبکه generator طیف کوچکی از داده ها را بیابد که به سادگی توان فریب discriminator را دارند. برخی از راه حل های پیشنهادی برای حل mode collapse به شرح زیر می باشد:

- استفاده از معماری W-GAN
- استفاده از روش Unrolling : به روز رسانی وزن های generator پس از k مرحله از به روز رسانی وزن های discriminator. اینکار باعث می شود شبکه generator تا چند مرحله از آینده را رصد کرده و سپس تشویق به تولید خروجی های متنوع تری شود.
- استفاده از روش Packing : ارتقا دادن discriminator به گونه ای که تصمیم خود را بر اساس چندین نمونه از یک کلاس برای تشخیص جعلی یا حقیقی بودن اتخاذ کند.

(ب) پاسخ این سوال به صورت مشروح در این لینک داده شده است. توجه شود پاسخی خلاصه بر مبنای این مقاله یا توضیحات داخل جزوه کاملا پذیرفته بوده و نمره کامل خواهد گرفت.

▷

مسئله‌ی ۵. (۵ + ۱۰ نمره)

(آ) یکی از چالش های مهم که مدل های GAN تاثیر به سزایی در برطرف شدن آن ها دارند موضوع image2image translation می باشد. یکی از مدل هایی که به طور خاص برای این امر توسعه داده شده است مدل Cycle GAN می باشد. معماری دوگانه این مدل را تشریح کرده و مزایای آن را در ارتباط با چالش مذکور نسبت به مدل پایه GAN بیان نمایید.

(ب) فرض کنید تحت شرایطی بسیار خاص از شما درخواست شده با تعداد داده محدود (در حدی که معماری پایه GAN به سختی به کمک آن train می شود) یک مدل Cycle GAN را train نمایید. در صورتی که الزام به استفاده از Cycle GAN باشد، راه حل پیشنهادی شما چیست؟ (دقت شود سوال ابتکاری بوده و پاسخ کاملا یکتا ندارد، اما بایستی هر راه حل پیشنهادی با تحلیل کامل و حتی استفاده از روابط ریاضی در صورت نیاز تشریح گردد).

حل.

(آ) پاسخ این سوال به صورت مشروح در این لینک داده شده است. توجه شود پاسخی خلاصه بر مبنای این مقاله کاملا پذیرفته بوده و نمره کامل خواهد گرفت.

▷

سوالات عملی (۵ + ۴۰ نمره)

مسئله‌ی ۶. (۲۰ نمره)

هدف این سوال طراحی یک شبکه ساده GAN می‌باشد که بتواند تصاویر دادگان MNIST را تولید نماید. فایل GAN.ipynb را براساس موارد خواسته شده تکمیل نمایید. دقت کنید که بخش هایی از نمره بستگی به نتایج بدست آمده دارند. باتوجه به نکات تمرین و مواردی که در کلاس عنوان شده، ساختار شبکه و تابع خطا و پارامترهای دیگر را طوری طراحی و انتخاب نمایید که در نهایت تصاویر باکیفیتی توسط Generator تولید شود و فرایند آموزش پایدار باشد. در نهایت فایل تکمیل شده به همراه نتایج را بعلاوه فایل پارامترهای شبکه Generator ای که آموزش داده‌اید به همراه دیگر بخش‌های تمرین ارسال نمایید.

مسئله‌ی ۷. (۵ + ۲۰ نمره)

Conditional VAE یکی از ورژن‌های modified شده VAE بوده که بر خلاف VAE کلاسیک، متغیرهای مورد نیاز را به صورت conditioned نسبت به برخی متغیرهای تصادفی تخمین می‌زند. در این تمرین هدف مقایسه خروجی این دو مدل بر روی مجموعه داده MNIST می‌باشد. لطفا کد هر دو روش پیاده‌سازی شده و خروجی آن‌ها از بعد میزان وضوح تصاویر تولید شده مقایسه گردد.

برای طراحی شبکه های VAE و CVAE ، به طور کلی محدودیت چندانی وجود ندارد اما پیشنهاد می شود برای قسمت Encoder ، سه لایه کانولوشن دو بعدی به ترتیب با ۱۶، ۳۲، و ۳۲ لایه به همراه MaxPool دو بعدی ۲ در ۲ پس از هرکدام طراحی شود. برای قسمت Decoder نیز دو لایه خطی به ترتیب با ۳۲ و ۶۴ لایه طراحی گردد. برای تابع هزینه لطفا از Binary Cross Entropy به همراه KL Divergence استفاده شود. برای Optimizer نیز از Adam استفاده گردد. مابقی پارامترها همانند mean می‌توانند به صورت customize شده انتخاب گردد و بسته به خروجی بهتر تغییر کند. برای راهنمایی بیشتر می‌توانید از کد موجود در این لینک استفاده کنید. لطفا کد را کپی نکرده و صرفا برای کمک و الهام‌گیری کد زنی خود از آن استفاده شود. توجه شود حتی ساختار پیشنهادی شبکه بسته به صلاح دید شخصی شما قابل تغییر بوده فقط توجه شود که نوشتن گزارش بخش عملی الزامی بوده و دارای نمره می‌باشد لذا حتما تمامی مراحل اعم از ساختار شبکه‌ها و پارامترها باید به طور کامل در گزارش توضیح داده شوند. برای پیاده‌سازی نیز تنها مجاز به استفاده از کتابخانه pytorch می‌باشید.

توضیحات کلی:

به وضوح مشورت و همفکری در حل سوالات هیچگونه ای مشکلی ایجاد نخواهد کرد اما جواب سوالات به هیچ عنوان نباید یکسان باشد. هر فرد باید جداگانه و Unique پاسخ‌های خود را بنویسد و در صورت شباهت بسیار زیاد نمره بین افرادی که پاسخ های بسیار مشابه دارند تقسیم خواهد گشت. در صورتی که کد شما در بخش عملی به خطا برخورد و خطا قابل برطرف کردن نبود، نمره شما بر اساس کیفیت کد به میزان قابل قبولی لحاظ خواهد گشت. لذا لطفا پاسخ سوال‌های عملی را خالی نگذارید. برای سوال عملی لطفا کد و گزارش هردو ارسال شوند. کمبود یکی از آن‌ها منجر به از دست رفتن نمره خواهد شد.

موفق باشید :)