

شبکه‌های عصبی و یادگیری عمیق

دکتر صفابخش



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

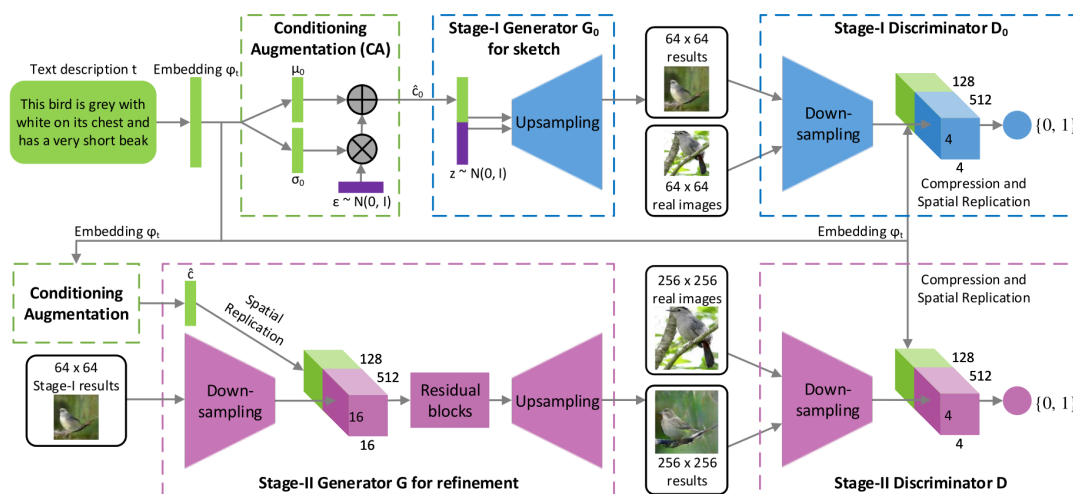
تمرین ششم
شبکه GAN

۱۴ تیر ۱۴۰۳

سوال اول - شبکه‌های مولد تقابلی

شبکه‌های مولد تقابلی^۱ همانطور که در کلاس با آنها آشنا شدید شامل دو زیرشبکه‌ی تولیدکننده^۲ و تمایزگر^۳ هستند که به صورت تقابلی آموزش داده می‌شوند تا داده‌های جدید تولید کنند. تولید جدید هدفی است که در تمامی مدل‌های مولد مد نظر قرار دارد و به شکل‌های مختلف از جمله ترجمه‌ی تصویر به تصویر، تبدیل دامنه و تولید شرطی صورت می‌گیرد. یکی از این اشکال، تولید تصویر با دریافت فرمان زبانی است که امروزه نیز نمونه‌های کاربردی آن همچون Imagen و Dall-E در دسترس عموم قرار دارند. در این تمرین به طور خاص به پیاده سازی این وظیفه با شبکه‌ی مولد تقابلی پشته‌ای یا StackGAN می‌پردازیم.

۱. به مراجعه به مقاله **StackGAN** کلیت ساختار و چگونگی عملکرد این شبکه را توضیح دهید. توضیح دهید که شبکه‌ی تعریف شده در هر گام^۴ به چه منظور استفاده می‌شود. به طور خاص ذکر کنید که ورودی شبکه‌ی تولیدکننده در هر دو گام چه تفاوتی با ورودی یک شبکه‌ی مولد تقابلی ساده^۵ دارد؟ همچنین بررسی کنید که آموزش این شبکه به چه صورت انجام می‌شود.



شکل ۱: معماری کلی شبکه مولد تقابلی پشته‌ای

Networks Adversarial Generative^۱
 Generator^۲
 Discriminator^۳
 Stage^۴
 Vanilla GAN^۵

پاسخ

همانطور که در توضیحات سوال بیان شد، تولید تصاویر با کیفیت بالا از متن، چالشی است که این مقاله آن را حل نموده است. این مقاله با شکستن کار به دو مرحله:

(a) Stage-I GAN

(b) Stage-II GAN

چالش تولید تصویر با وضوح و دقت بالا از داده‌های متنی را حل می‌کند.

(آ) مرحله اول:

• هدف: مرحله اول GAN مسئول تولید یک تصویر (64x64) با وضوح پایین ست که شکل و رنگ اساسی شیء توصیف شده توسط متن را ترسیم می‌کند.

• ورودی:

- بردار تبدیل شده متن: توضیحات متنی با استفاده از یک انکودر از پیش آموزش داده شده ϕ_t به یک بردار عددی تبدیل می‌شوند.

- متغیرهای شرطی گوسی: متغیرهای شرطی از توزیع گوسی که توسط تعبیه متنی پارامتره شده‌اند، نمونه‌برداری می‌شوند.

متغیرهای شرطی گوسی \hat{c}_0 برای تعبیه متنی از $N(\mu_0(\phi_t), \Sigma_0(\phi_t))$ نمونه‌برداری می‌شوند تا معنای ϕ_t را با تغییرات مختلف به تصویر بکشند. این می‌تواند پایداری آموزش و تنوع نمونه را فراهم کند، به عنوان مثال: انواع مختلف حالات و ظاهرها.

- بردار نویز: یک بردار نویز تصادفی z با متغیرهای شرطی ترکیب می‌شود.

• خروجی: یک تصویر با وضوح پایین که شکل اصلی و رنگ‌های اساسی متناظر با توضیحات متنی را به تصویر می‌کشد.

مشروط بر \hat{c}_0 و متغیر تصادفی z ، مرحله اول GAN متمایزکننده D_0 و تولیدکننده G_0 را با جایگزینی حداکثر کردن L_{D_0} در معادله (۱) و حداقل کردن L_{G_0} در معادله (۲) آموزش می‌دهد:

$$(1) \quad L_{D_0} = \mathbb{E}_{(I_0, t) \sim p_{data}} [\log D_0(I_0, \phi_t)] + \mathbb{E}_{z \sim p_z, t \sim p_{data}} [\log(1 - D_0(G_0(z, \hat{c}_0), \phi_t))],$$

$$(2) \quad L_{G_0} = \mathbb{E}_{z \sim p_z, t \sim p_{data}} [\log(1 - D_0(G_0(z, \hat{c}_0), \phi_t))] + \lambda D_{KL}(N(\mu_0(\phi_t), \Sigma_0(\phi_t)) \| N(0, I)),$$

تصویر واقعی I_0 و توضیح متنی t از توزیع داده‌های واقعی p_{data} هستند.

(ب) مرحله دوم:

• هدف: مرحله دوم GAN تصویر با وضوح پایین تولید شده در مرحله اول را تصحیح می‌کند تا یک تصویر (256x256) با وضوح بالا با جزئیات بیشتر و واقع‌گرایانه‌تر تولید کند.

پاسخ

• ورودی:

- خروجی مرحله اول: تصویر با وضوح پایین از مرحله اول.
- تعبیه متنی: تعبیه متنی دوباره برای شرطی‌سازی فرآیند تولید استفاده می‌شود، که اطمینان حاصل کند جزئیات تصویر با متن همخوانی دارد یا خیر.
- متغیرهای شرطی گوسی: متغیرهای شرطی جدید از توزیع گوسی که توسط تعبیه متنی پارامتره شده‌اند، نمونه‌برداری می‌شوند.
- خروجی: یک تصویر با وضوح بالا و جزئیات دقیق که واقع‌گرایانه است که توضیحات متنی را به درستی منعکس می‌کند.

$$L_D = \mathbb{E}_{(I,t) \sim p_{data}} [\log D(I, \phi_t)] + \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}} [\log(1 - D(G(s_0, \hat{c}), \phi_t))], \quad (3)$$

$$L_G = \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}} [\log(1 - D(G(s_0, \hat{c}), \phi_t))] + \lambda D_{KL}(N(\mu(\phi_t), \Sigma(\phi_t)) \| N(0, I)), \quad (4)$$

مرحله دوم GAN یاد می‌گیرد تا اطلاعات مفید در تعبیه متنی که توسط مرحله اول GAN نادیده گرفته شده است را به دست آورد.

- تولیدکننده مرحله دوم یک شبکه انکودر-دکودر با بلوک‌های residual است.
- ویژگی‌های تصویر و ویژگی‌های متن در طول بعد کانال ترکیب می‌شوند تا نمایش‌های چندوجهی در سراسر ویژگی‌های تصویر و متن را یاد بگیرند.
- در نهایت، از یک سری لایه‌های دیکودر برای تولید یک تصویر با وضوح بالا $W \times H$ استفاده می‌شود.
- به جای استفاده از متمایزکننده ساده، از متمایزکننده matching-aware برای هر دو مرحله استفاده می‌شود.

تفاوت‌های ورودی در مقایسه با GAN های ساده:

(آ) GAN های ساده:

- ورودی: یک بردار نویز ساده به عنوان ورودی برای شبکه تولیدکننده استفاده می‌شود.
- خروجی: یک فرآیند تولید تک‌مرحله‌ای که به طور مستقیم سعی در ایجاد تصویر با وضوح بالا دارد.

StackGAN: (ب)

- ورودی مرحله اول: ترکیبی از بردار نویز و متغیرهای شرطی مشتق شده از تعبیه‌های متنی برای تولید یک تصویر با وضوح پایین.
- ورودی مرحله دوم: از خروجی مرحله اول استفاده می‌کند و تعبیه‌های متنی را با متغیرهای شرطی جدید برای تصحیح تصویر و افزودن جزئیات مجدداً استفاده می‌کند.

۲. شبکه‌های مولد تقابلی در مقایسه با سایر شبکه‌ها از سه مشکل اساسی رنج می‌برند: این سه مشکل عبارتند از فروپاشی مد، عدم همگرایی و ناپدید شدن گرادیان. به طور مختصر توضیح دهید که هر کدام به چه صورتی و چه راهکارهایی برای رفع آنها مطرح شده است؟

پاسخ

(آ) فروپاشی مد:

فروپاشی مد زمانی رخ می‌دهد که مولد تعداد محدودی خروجی تولید می‌کند و عملاً به چند مد (یا حتی یک مد) از توزیع داده‌ها فرو می‌ریزد. این به این معناست که GAN نمی‌تواند تمام تنوع توزیع داده‌های واقعی را به تصویر بکشد. برای حل این مشکل می‌توان از موارد زیر استفاده نمود:

- **تمایز دسته‌ای (Minibatch Discrimination):** این تکنیک به تشخیص‌دهنده اجازه می‌دهد تا نمونه‌های متعدد را با هم بررسی و مقایسه کند، و مولد را تشویق می‌کند تا خروجی‌های متنوع‌تری تولید کند.
- **GAN های بدون بازگشت (Unrolled GANs):** این رویکرد فرآیند آموزش را تغییر می‌دهد تا بهینه‌سازی تشخیص‌دهنده را Unrolled کند و اثر به‌روزرسانی‌های آینده را در نظر بگیرد که به جلوگیری از فروپاشی مد کمک می‌کند.
- **مجموعه‌ای از مولدها (Ensemble of Generators):** استفاده از مولدهای متعدد می‌تواند به پوشش مدهای مختلف توزیع داده‌ها کمک کند و بنابراین فروپاشی مد را کاهش دهد.

(ب) عدم همگرایی:

عدم همگرایی یا همگرایی ناکامل زمانی رخ می‌دهد که فرآیند آموزش GAN نوسان کند یا نتواند به یک حالت پایدار برسد. این ناپایداری ناشی از این است که مولد و تشخیص‌دهنده دائماً سعی در شکست دادن یکدیگر دارند. برای اینکه شبکه ناپایدار نشود می‌توان راهکارهای زیر را انجام داد:

- **معماری‌های بهبود یافته:** استفاده از تکنیک‌هایی مانند نرمال‌سازی طیفی که با کنترل ثابت Lipschitz تشخیص‌دهنده، آموزش را پایدار می‌کند.
- **توابع هدف اصلاح شده:** توابع زیان جایگزین، مانند زیان Wasserstein (استفاده شده در Wasserstein GANs یا WGANs)، می‌تواند همگرایی را با ارائه یک گرادیان معنادارتر برای مولد حتی زمانی که تشخیص‌دهنده نزدیک به بهینه است، بهبود بخشد.
- **تکنیک‌های تنظیم (Regularization):** تکنیک‌هایی مانند gradient penalty (استفاده شده در WGAN-GP) می‌تواند با جریمه کردن نرم گرادیان تشخیص‌دهنده، آموزش را پایدار تر کند و از ناپایداری جلوگیری کند.

(ج) محو شدن گرادیان:

ناپدید شدن گرادیان زمانی رخ می‌دهد که گرادیان‌های بازگشتی از طریق شبکه بسیار کوچک می‌شوند و منجر به یادگیری آهسته یا عدم یادگیری شبکه مولد می‌شوند. این مشکل زمانی که تشخیص‌دهنده بسیار قوی شود و به راحتی بین داده‌های واقعی و جعلی تمایز قائل شود، بیشتر می‌شود. برای قلبه به این مشکل راه‌حل‌های زیر پیشنهاد می‌شود:

Mode Collapse^۶

پاسخ

- استفاده از WGAN: چارچوب Wasserstein GAN مشکل ناپدید شدن گرادیان‌ها را با اطمینان از اینکه گرادیان‌ها ناپدید نمی‌شوند، حل می‌کند. این کار را با استفاده از تابع زیان متفاوت بر اساس فاصله Wasserstein انجام می‌دهد که سیگنال گرادیان پایدارتر و مفیدتری ارائه می‌دهد.
- تطبیق ویژگی (Feature Matching): این تکنیک شامل آموزش مولد برای تطبیق آمار ویژگی‌های استخراج شده از یک لایه واسطه تشخیص‌دهنده است، به جای اینکه به طور مستقیم سعی کند تشخیص‌دهنده را فریب دهد.
- تعادل در آموزش (Training Balance): اطمینان از اینکه مولد و تشخیص‌دهنده ظرفیت‌های یادگیری متعادلی دارند و به طور متعادل آموزش داده می‌شوند می‌تواند از قوی شدن بیش از حد تشخیص‌دهنده نسبت به مولد جلوگیری کند و یادگیری هردو شبکه را در یک حد نگه دارد.

۳. یک ایده‌ی رایج برای بهبود عملکرد شبکه‌های مولد تقابلی استفاده از عملگر PixelShuffle است. نحوه‌ی عملکرد این عملگر و تأثیر آن را بررسی کنید. بررسی کنید که این عملگر اولین بار در چه وظیفه‌ای و به چه منظور تعریف شد؟ همچنین بررسی کنید که به طور خاص در معماری StackGAN در کدام زیرشبکه‌ها قابل استفاده است و چه عملکردی خواهد داشت؟

پاسخ

عملگر PixelShuffle که به عنوان هم‌ترازی زیرپیکسلی نیز شناخته می‌شود، تکنیکی برای بهبود وضوح مکانی تصویر است. این عملگر عناصر یک تانسور را به گونه‌ای بازآرایی می‌کند که یک تانسور با وضوح مکانی بالاتر و عمق کانال کمتر تولید شود. این عملگر یک تانسور ورودی با شکل $(B, C \times r^2, H, W)$ را دریافت می‌کند و آن را به تانسوری با شکل $(B, C, H \times r, W \times r)$ تبدیل می‌کند، که در آن B اندازه دسته، C تعداد کانال‌ها، H و W ارتفاع و عرض تصویر و r ضریب بزرگ‌نمایی است. روابط ریاضی آن را می‌توان به صورت زیر بیان نمود:

با توجه به تانسور ورودی X با شکل $(B, C \times r^2, H, W)$ ، عملگر PixelShuffle آن را به تانسور خروجی Y با شکل $(B, C, H \times r, W \times r)$ بازآرایی می‌کند:

$$Y(b, c, h \times r + i, w \times r + j) = X(b, c \times r^2 + i \times r + j, h, w) \quad (5)$$

برای بازه‌های:

- (a) $b \in [0, B)$
- (b) $c \in [0, C)$
- (c) $h \in [0, H)$
- (d) $w \in [0, W)$
- (e) $i, j \in [0, r)$

عملگر PixelShuffle تأثیرات مفید زیادی بر GAN‌ها دارد، به‌ویژه در وظایفی که نیاز به تولید تصاویر با وضوح بالا دارند برای مثال می‌توان به موارد زیر اشاره نمود:

پاسخ

(آ) وضوح بهبود یافته: با بازآرایی موثر عناصر تنسور، PixelShuffle وضوح تصویر را افزایش می‌دهد که برای تولید تصاویر با کیفیت بالا حیاتی است.

(ب) کارایی: PixelShuffle از لحاظ محاسباتی کارا است زیرا نیاز به لایه‌های بزرگ‌نمایی و کانولوشن‌های اضافی را کاهش می‌دهد و شبکه را سریع‌تر و با منابع کمتر می‌سازد.

(ج) استفاده بهتر از ویژگی‌ها: این عملگر به استفاده بهتر از ویژگی‌های یادگرفته شده توسط شبکه کمک می‌کند زیرا مستقیماً وضوح مکانی را بدون از دست دادن جزئیات یادگرفته شده بهبود می‌بخشد.

عملگر PixelShuffle ابتدا در زمینه وظایف سوپررزولوشن، به‌ویژه در مقاله Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network (ESPCN) توسط Wenzhe Shi و همکاران تعریف شد. هدف این بود که با یادگیری یک عملگر کانولوشن زیرپیکسلی کارا، وضوح تصویر را بدون هزینه محاسباتی قابل توجهی افزایش دهد. استفاده در معماری StackGAN در زیرشبکه دوم:

(آ) در مرحله دوم معماری PixelShuffle هدف بهبود رزولوشن تصاویر تولید شده در مرحله اول است حال اگر با عملگر PixelShuffle ادغام شود، مدل می‌تواند تصاویر با جزئیات بیشتر و کیفیت بالاتری تولید کند.


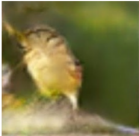











(ب) PixelShuffle پیچیدگی محاسباتی عملیات بزرگ‌نمایی را کاهش می‌دهد که منجر به زمان‌های آموزش سریع‌تر می‌شود.

۴. معیار FID (Frechet Inception Score) یک معیار برای ارزیابی کیفیت و تنوع تصاویر تولید شده توسط مدل‌های مولد است. توضیح دهید که این معیار به چه صورت محاسبه می‌شود، به چه ویژگی‌هایی از مدل و یا داده وابسته است و آیا معیار قابل اتکایی برای مقایسه‌ی مدل‌های مولد محسوب می‌شود؟

پاسخ

۵. مدل را بر روی این داده‌ها آموزش دهید. معماری نهایی هر یک از چهار زیرشبکه به همراه نمودار خطای تولیدکننده و تمایزگر در هر گام آموزش را در گزارش خود بیاورید. پس از پایان آموزش ۱۰ تصویر را به صورت تصادفی از خروجی مدل در stage اول و دوم تولید کنید.

برای این پروژه از مجموعه داده‌ی CUB-2011 استفاده می‌کنیم که شامل یازده هزار تصویر از ۲۰۰ گونه پرنده می‌باشد و به ازای هر تصویر یک توصیف متنی نیز وجود دارد. **مجموعه‌ی داده** در سایت Kaggle و توصیفات متنی نیز در **این لینک** موجود است. همچنین برای توصیفات متنی از پیش تعبیه‌ی آماده شده در فایل char-CNN-RNN-embeddings.pickle وجود دارد که می‌تواند جایگزین ساختن داده باشد. استفاده از پیش تعبیه‌ها نیز که منجر به کارایی بهتر مدل شوند دارای ۵ امتیاز اضافی می‌باشد.

Text description	This bird is blue with white and has a very short beak	This bird has wings that are brown and has a yellow belly	A white bird with a black crown and yellow beak	This bird is white, black, and brown in color, with a brown beak	The bird has small beak, with reddish brown crown and gray belly	This is a small, black bird with a white breast and white on the wingbars.	This bird is white black and yellow in color, with a short black beak
Stage-I images							
Stage-II images							

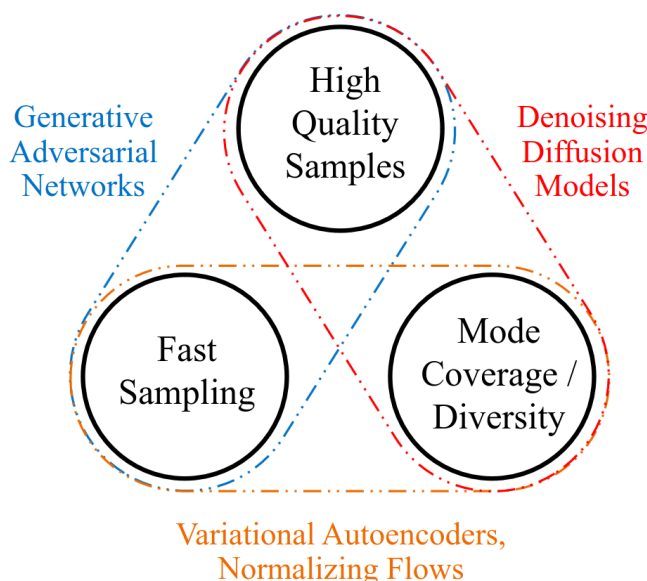
شکل ۲: نمونه خروجی مدل StackGAN برای مجموعه داده CUB

پاسخ

سوال دوم - مدل‌های پخشی (امتیازی)

مدل‌های مولد حوزه‌ی تصویر به چهارچوب‌های مختلف تقسیم می‌شوند. وجه مشترک همه‌ی این مدل‌ها این است که تلاش می‌کنند تا با یادگیری توزیع داده‌ها نمونه‌های جدیدی از آن تولید کنند. تاکنون مدل‌های مولد حوزه‌ی تصویر را می‌توان در چهار قالب کلی شامل خودکدگذارهای تغییراتی^۷، شبکه‌های مولد تقابلی، جریان‌های نرمال‌ساز^۸ و مدل‌های پخشی دسته‌بندی کرد که در هر قالب انواع مختلفی از پیاده‌سازی‌ها وجود دارد.

۱. در این مقاله سه نیازمندی کلی برای کارایی یک مدل مولد حوزه‌ی تصویر ذکر می‌شود که عبارتند از: تولید نمونه‌های با کیفیت، سرعت بالای تولید نمونه و تنوع نمونه‌های تولیدی. و نیز اشاره می‌شود که هر مدل مولدی که تاکنون در یکی از قالب‌هایی که بالاتر ذکر شد ارائه شده است در یکی از این سه نیازمندی ضعیف عمل می‌کند. با بررسی مقاله توضیح دهید که هر مدل در چه زمینه‌ای و به چه علتی ضعیف عمل می‌کند؟



شکل ۳: مشکل سه‌گانه مدل‌های مولد

مدل‌های پخشی دسته‌ای از مدل‌های مولد هستند که در حال حاضر بهترین مدل تولید تصویر شناخته می‌شوند. در مدل‌های پخشی احتمالاتی از یک ذخیره‌ی مارکف برای مدل کردن فرآیند نویززدایی و نویز افزایشی استفاده می‌شود و دو مسیر کلی رو به جلو^۹ و رو به عقب^{۱۰} در نظر گرفته می‌شود. در مسیر رو به جلو داده‌ی اولیه مرحله به مرحله با نویز تخریب می‌شود تا به یک نویز گاوسی تبدیل شود و در فرآیند رو به عقب نیز نویز زدایی با شروع از یک نویز تصادفی اولیه انجام می‌شود تا به نمونه‌ای جدید از توزیع داده‌ها برسیم. این موضوع در تصویر «۴» آورده شده است.



شکل ۴: فرایند کلی مدل‌های پخشی

Diffusion Models^۷
Flows Normalizing^۸
Forward^۹
Backward^{۱۰}

۱. با توجه به مقاله‌ی **مدل‌های پخشی احتمالاتی**، در مسیر رو به جلو نیازی به اضافه کردن نویز به صورت مرحله به مرحله نیست و می‌توان نویز اضافه شونده به هر مرحله را به صورت مستقیم و با استفاده از رابطه‌ی زیر به دست آورد:

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

با استفاده از خاصیت نویز گاوسی رابطه‌ی بالا را اثبات کنید. (راهنمایی: اگر دو متغیر نرمال مستقل داشته باشیم جمع آنها نیز نرمال است.)

۲. با توجه به مقاله‌ی سوال قبل، فرآیند آموزش و نمونه برداری مدل‌های پخشی را توضیح دهید. در فرآیند رو به عقب یک فرض مهم این است که توزیع $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ گاوسی است. در چه صورتی این فرض درست است؟

۳. یک مسالهی اساسی در مدل‌های پخشی این است که در هیچ یک از گام‌ها محاسبات در بعد کوچکتري صورت نمی‌گیرند در نتیجه در صورت بزرگ بودن دیتاست و اندازه‌ی تصاویر ورودی مدل‌های پخشی بسیار پرهزینه و حجیم خواهد شد. مدل Latent Stable Diffusion برای حل این چالش از چه رویکردی استفاده می‌کند؟ به نظر شما چرا برای کاهش حجم نیاز به مدل‌های تغییراتی^{۱۱} داریم؟

۴. یک مدل پخشی را بر روی داده‌های MNIST آموزش دهید. تعداد پارامترهای مدل، تصاویر میان فرآیند رو به جلو و فرآیند رو به عقب را برای یک تصویر گزارش کنید. همچنین ۵ تصویر تولید شده توسط این مدل را نمایش دهید.