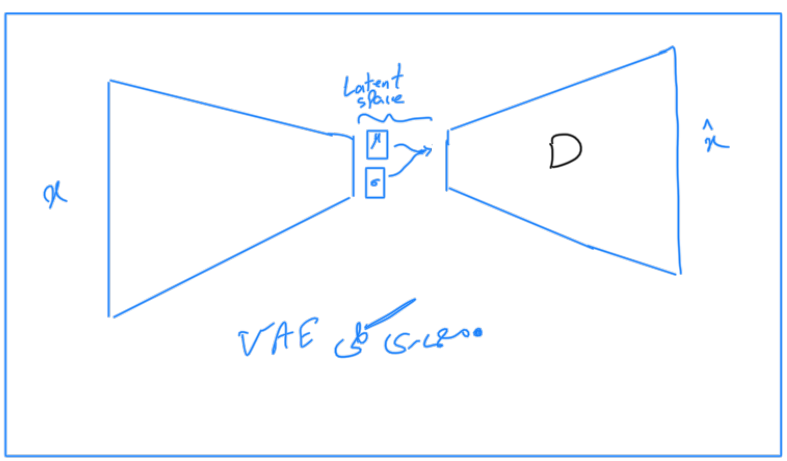


VAE



VAE یک مدل Generative هست که سعی در ساخت یک latent space معنی دار از روی توزیع دیتا با استفاده از encoder و بازپسین تصویر از این فضای پتان مناسب تصویر آموزش با استفاده از Decoder دارد.

$$P_\theta(x) = \int_z P_\theta(x|z) P_\theta(z) dz$$

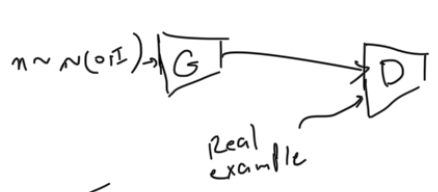
در زمان آموزش برای اینکه تصویر تولیدی شبیه تصویر ورودی باشد، از تابع هزینه $E_{z \sim q_\phi(z|x)} \|x - D_\theta(z)\|^2$ استفاده می کنند:

$$-\frac{1}{2} E_{z \sim q_\phi(z|x)} [\|x - D_\theta(z)\|^2]$$

و برای حفظ ویژگی های توزیع نهان، overfit شدن این فضا به توزیع دیتای آموزشی از تابع هزینه KLD استفاده می کنند:

$$L_{\theta, \phi} = -\frac{1}{2} E_{z \sim q_\phi(z|x)} [\|x - D_\theta(z)\|^2] + D_{KL}(q_\phi(z|x) || P_\theta(z))$$

GAN



GAN به صورت مقابل است:

در این مدل سعی در ساخت نمونه های مشابه دیتای آموزشی با استفاده از Generator داریم و برای اکتزایی کیفیت نمونه ها از یک Discriminator جهت تشخیص نمونه های تولیدی generator از داده های واقعی است. در هر چرخ یک مسابقه بین Generator و Discriminator وجود دارد.

$$L_D = -E_{x \sim P_{data}} [\ln D(x)] - E_{x \sim G} [\ln (1 - D(x))]$$

$$L_G = -E_{x \sim G} [\ln D(x)]$$

Diffusion

در این مدل ها، در فرآیند forward به صورت پله ای تصویر را نویز می کنند و در فرآیند backward پله ای تصویر را denoise می کنند تا تصویر حس کننده به تصویر در فضای تصاویر بدست آید. لازم به ذکر است فرآیند iterative است و ردی فضای خود تصویر را حس کنند نه فضای نهان و هم چنین تصویر نهایی تولید شده در فرآیند backward حس تواند به تصویر اولیه در فرآیند forward متفاوت باشد.

معمولاً برای Denoising از شبکه های U-net استفاده می کنند.

معمولاً تصاویر خروجی VAE متنوع است اما از نظر کیفیت با sharpness مشکل دارد و ریزش دارد. β VR

GAN ریزش بالایی دارند اما از نظر تنوع و رعایت جزئیات خوب است. β VR

Diffusion ریزش و تنوع بسیار خوبی دارند.

کاربرد VAE : Data Compression

GAN : Image-generation, style transfer, image-to-image translation

Diffusion : Denoising, image generation, ...

(b) در حوزه تحلیل تصاویر پزشکی مشکلات دیتاست کوچک، ریزش بودن نمونه‌ها رایج است. مدل‌های دینسی‌تر با توجه به ماهیتشان می‌توانند در مواجه با دیتاست کوچک و ریزش کم‌ترد خوبی از خود نشان دهند. هم چنین در این حوزه به دلیل مدل‌های متفاوت و شرایط مختلف تصویربرداری نیاز به uncertainty handling هست که مدل‌های دینسی‌تر با فکر کمتری دارند.

(2)

(a) با داشتن دیتای آموزشی یکسری نمونه داخل فضای کلی داریم برای رسیدن از یک نقطه ریزم داخل مقادیر فضای معنی‌دار دیتای آموزشی نیاز به تخمین گزیده‌ای از توزیع داده‌ها داریم. $(p_{\theta}(x))$ و با استفاده از مدل بهداری Generation را انجام دهیم. با توجه به اینکه دیتای آموزشی یک زیرمجموعه بسیار کوچکی از فضای کلی را تشکیل می‌دهند اکثر نقاط ریزم در فضای خروجی گمراهه را با استفاده از تابع امتیاز در هر مرحله سعی در تولید تصویر داشته باشیم، گزیده‌ای‌های بهارتری به دست نمی‌آوریم زیرا این نتایج ناشی از تصادف در تصاویر گمراهه‌های Invalid می‌تواند باشد. Generation را به سمت توزیع اصلی داده‌ها آموزش نمی‌توانیم و باید به سمت تولید با حذف کردن نمونه Data manifold را به سمت گزیده‌ای‌های بهارتری از فضای اصلی کشانیم تا بتوانیم گزیده‌ای‌های معنی‌دار را تولید کنیم و در هر مرحله که به manifold اصلی نزدیک می‌شویم می‌توانیم میدان این نمونه را کاهش داد.

$$p(x, t) = \sum_{j=1}^M w_j \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma^2 t)}} \exp \left\{ -\frac{(x - \mu_j)^2}{2(\sigma_j^2 + \sigma^2 t)} \right\} \quad (b)$$

$$\nabla_x \log p(x, t) = \frac{1}{p(x, t)} \cdot \nabla p(x, t)$$

$$\nabla p(x, t) = \sum_{j=1}^M w_j \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma^2 t)}} \left(\frac{-2(x - \mu_j)}{2(\sigma_j^2 + \sigma^2 t)} \right) \exp \left\{ -\frac{(x - \mu_j)^2}{2(\sigma_j^2 + \sigma^2 t)} \right\}$$

$$\sum_{j=1}^M w_j \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma^2 t)}} \left(\frac{-2(x - \mu_j)}{2(\sigma_j^2 + \sigma^2 t)} \right) \exp \left\{ -\frac{(x - \mu_j)^2}{2(\sigma_j^2 + \sigma^2 t)} \right\}$$

$$\Rightarrow \nabla_x \log p(x, t) = \frac{\sum_{j=1}^M w_j \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma^2 t)}} \exp \left\{ -\frac{(x - \mu_j)^2}{2(\sigma_j^2 + \sigma^2 t)} \right\}}{\sum_{j=1}^M w_j \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma^2 t)}} \exp \left\{ -\frac{(x - \mu_j)^2}{2(\sigma_j^2 + \sigma^2 t)} \right\}}$$

$$S(x, t) \Big|_{t=0} = \frac{\sum_{j=1}^M w_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(x - \mu_j)^2}{2(\sigma_j^2)} \right\} \left(-\frac{x - \mu_j}{\sigma_j^2} \right)}{\sum_{j=1}^M w_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(x - \mu_j)^2}{2(\sigma_j^2)} \right\}}$$

(3)

(a) فرض کنیم در دسته‌بندی تصایر CT مربوط به برای تشخیص Covid داشتن یا نداشتن، برای مجموعه dev set فریم‌های که 5 درصد خطا مربوط به کلاس positive و 95 درصد مربوط به کلاس negative باشد. از این درصد‌ها می‌توانیم که تمرکز اصلی برای بهبود عملکرد مدل به‌دلیل برتری کلاس negative باشد و اگر بخواهیم خطاهای این دسته را رفع کنیم عملکرد مدل بسیار بهبود خواهد یافت. همچنین در تحلیل تصایر مجموعه تست می‌بینیم که برخی از آنها از نظر توزیع با توزیع موجود که آموزش متفاوت هستند و نیاز به augmentation و یا Pre Process های مخصوص مانند noise reduction و blur دارند. به همین سبب تحلیل خطا باعث می‌گردد نف و منصف دقت مدل را بهبود دهد و برای حل آنها از سیدهای درستی استفاده کنیم.

(b) فرض کنیم برای تحلیل خطا، 5000 نمونه dev set باشد که بر 20 آن دای خطا هست. آنالیز 1000 تصویر فرایند بسیار زمان‌بر است برای این کار dev set را به درجهت eyeball که به آن نگاه کرده و تحلیل خطا می‌کنیم (این کار black box که برای آن نگاه نمی‌کنیم تقسیم می‌کنیم. از سمت eyeball برای تحلیل خطا و از سمت black box برای تنظیم hyperparameters و ارزیابی مدل استفاده و آنالیز کردن نرخ خطا استفاده می‌گردد. اگر در هنگام ارزیابی عملکرد eyeball بسیار بهتر از black box شود بین معنی است که overfitting به‌دلیل می‌گردد eyeball نرخ داده است و باید میزان دیتای eyeball بیشتر کرده و باز داده‌های به‌صورت دوره‌ای استفاده کرد.

(c) فرض کنیم برای حل تشخیص HER2 در breast cancer از یک پایلوت استفاده می‌کنیم nuclei ها و سپس طبقه‌بندی با استفاده از CNN ها استفاده کنیم. برای بهبود سیستم باید بدانیم که کدام قسمت از پایلوت منجر به خطای بیشتر می‌گردد به عنوان مثال اگر از 100 خطا، 95 خطا مربوط به استفاده از nuclei ها باشد و فقط 5 درصد مربوط به خطای دسته‌بندی CNN، برای بهبود مدل باید تمرکز به بهبود استفاده از nuclei ها باشد تا بتوان خطا را به خوبی کاهش داد.

(4)

$$J_{naive}(\theta) = \frac{1}{2} \int P(m, t) \left[S_{\theta}(m, t) - \nabla_x \log P(m, t) \right]^2 dx dt = \frac{1}{2} E_{P(m, t)} \left[\| S_{\theta}(m, t) - \nabla_x \log P(m, t) \|^2 \right]$$

$$= \frac{1}{2} \left(\underbrace{E_{P(x, t)} [\| S_{\theta}(x, t) \|^2]}_{J_{naive}} + \underbrace{E_{P(x, t)} [\| \nabla_x \log P(x, t) \|^2]}_{J_{model}} + \underbrace{E_{P(x, t)} [\langle S_{\theta}(x, t), \nabla_x \log P(x, t) \rangle]}_{Z(\theta)} \right)$$

عبارت مشترک در J_{naive}, J_{model}

$$Z(\theta) = \int P(\cdot) \langle S_{\theta}(\cdot), \nabla_x \log P(\cdot) \rangle d\cdot$$

log derivative trick

$$= \int P(\cdot) \langle S_{\theta}(\cdot), \frac{\nabla_x P(\cdot)}{P(\cdot)} \rangle d\cdot = \int \langle S_{\theta}(\cdot), \nabla_x P(\cdot) \rangle d\cdot$$

Bayesian Inference

$$= \int \langle S_{\theta}(\cdot), \nabla_x \int P(x^0) P(\cdot | x^0) dx^0 \rangle d\cdot = \int \langle S_{\theta}(\cdot), \int P(x^0) \nabla_x P(\cdot | x^0) dx^0 \rangle d\cdot$$

log derivative trick

$$= \int \langle S_{\theta}(\cdot), \int P(x^0) P(\cdot | x^0) \nabla_x \log P(\cdot | x^0) dx^0 \rangle d\cdot = \int P(x^0) P(\cdot | x^0) \langle S_{\theta}(\cdot), \nabla_x \log P(\cdot | x^0) \rangle dx^0 d\cdot$$

در نتیجه داریم:

J_{naive} و J_{model} کمینه‌ها یکسان اند.

$J_{naive}(\theta) = J_{model}(\theta) + const1 - const2 \Rightarrow$

یکسان عبارت
ثابت نسبت به θ
متفاوت مستند