

(11) استیلاچ L خودمان را بعضی تغییرات جای کشیم E_f ، h هسته پس همان طور که می دانیم اولین کارمه

coordinate ascent
variational inference

قابل حساب غیر قابل حساب

$$\nabla_{\theta} L = \nabla_{\theta} E_q[f(z)] + \nabla_{\theta} h_{\theta}(x) \Rightarrow \nabla_{\theta} E_q[f(z)] = \nabla_{\theta} \int_z f(z) q_{\theta}(z) dz$$

$$\int_z f(z) \nabla_{\theta} q_{\theta}(z) dz = \int_z f(z) q_{\theta}(z) \nabla_{\theta} \ln q_{\theta}(z) dz \Rightarrow \nabla_{\theta} E_q[f(z)] = E_q[f(z) \nabla_{\theta} \ln q_{\theta}(z)]$$

البته از رابطه $\nabla_{\theta} q_{\theta}(z) = q_{\theta}(z) \nabla_{\theta} \ln q_{\theta}$ نیز استفاده کردیم.

حال برای تخمین stochastic داریم

$$\nabla_{\theta} E_q[f(z)] = E_q[f(z) \nabla_{\theta} \ln q_{\theta}(z)] \approx \frac{1}{N} \sum_{i=1}^N f(z^i) \nabla_{\theta} \ln q_{\theta}(z^i)$$

همان طور که میدانیم (1) (2) (3) داده VAE دو نوع نوی AE است که توزیع آنکوینف جاده مرکزی آموزش به نحوی به سبب سببای می شود که فضای

پنهان (latent) آن ویژگی هایی را پیدا کند که بتواند این به تئوری تکیه کند و از $variational$ از استقامت از $variational$ inference

می آید پس به جای مدل سازی و بررسی به عنوان یک نقطه ، ما آن را به عنوان توزیع از فضای پنهان مدل سازی می کنیم

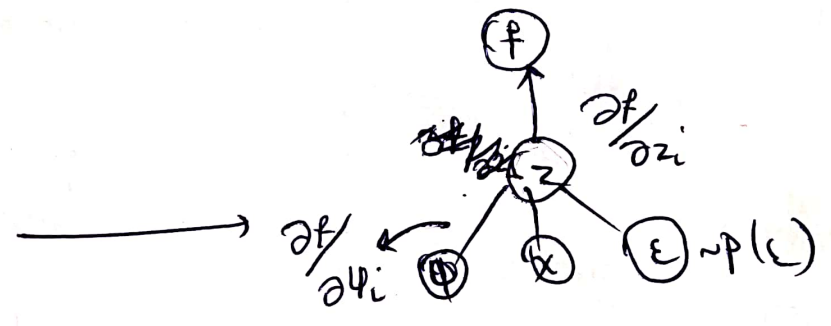
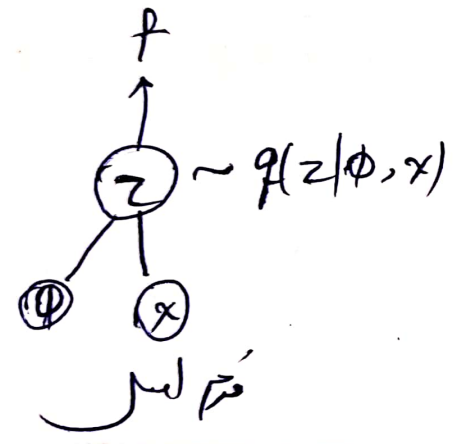
پس دو نوع داریم بهترین توزیع هدف را تشخیص می دهیم که به این معنای توان از استخراج معنی است

۱) ادله این سخن نا اریب (unbiased) است ولی واریانس بالا دارد که به سادگی قابل حل است.
(مگر اعدادی دیا که آن هم همیشه محدود است)

(الف) این روش برای این به کار می رود که بتوانیم از sampling به جای ~~محاسبه~~ محاسبه مستقیم راداسته باشیم

می دانیم مشتق گرفتن از انتان تصادفی انتان پذیر نیست ولی با روش پارامتریزه کردن انتان مشتق می گیریم. دواتر VAE از:

فرد z ~~انتان پذیر~~ انتان پذیری کننده این نویز برداری با متغیر پارامتریک $q(z|\phi, x)$ تعیین زده می شود. ~~backprop~~ backprop حال یک پارامتر جدید
 و تعریف می شود که z را به نویز تبدیل کند. اجاره backprop جدید

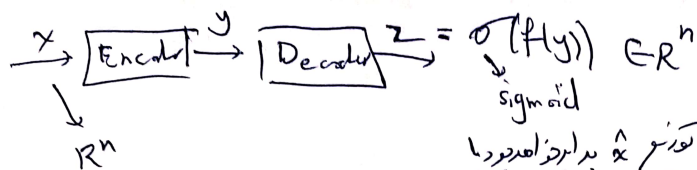


عملاً به جای این که یک تابع نویز داشته باشیم یک توزیع نرمال استاندارد داریم

① ج. اگر توابع گوی باشد به شرط این که در واریش بعد (تقریباً) یک باشد و ثابتاً $\frac{1}{2} \frac{d^2 \log L}{d\theta^2} < 0$ باشد.

کواریش بعد صحت باشد یعنی ماتریس کواریش $[Cov(\hat{\theta})]$ باشد $\frac{1}{2} \frac{d^2 \log L}{d\theta^2} < 0$ باشد و باید از این

MSR $\frac{1}{2} \frac{d^2 \log L}{d\theta^2} < 0$ باشد (البته در منابع گفته شده که همیشه باید باشد ولی با تقدم تقریباً همیشه این شرط برقرار است)
 $\frac{1}{2} \frac{d^2 \log L}{d\theta^2} < 0$



(۲) الف

حال \hat{x} با Sampling از z ساختی و سپس توزیع \hat{x} بدیفرانسیال می‌گیری

$$P(\hat{x}=x|y) = \prod_{i=1}^n z_i^{x_i} (1-z_i)^{(1-x_i)} \Rightarrow \log P(\hat{x}=x|y) = \sum_{i=1}^n x_i \log z_i + (1-x_i) \log(1-z_i)$$

↓

که این همان Binary Cross Entropy حقیقت است.

(۲) ب راه حل این سنده Gumbel Softmax است. به طور خلاصه یک روش تقریبی ساده برای داده‌های

categorical است که می‌تواند با استفاده از backprop آگرادینت را با استفاده از (reparametrization)

توزیع تصادفی آن به این صورت است که بتوانیم کاری کنیم که احتمال‌های گسسته و احتمالاتی حذف شود

$$z \sim \text{Cat}(\pi_1, \dots, \pi_n)$$

یعنی z یک متغیر تصادفی categorical باشد

احتمال‌های π_i که منجر به z می‌شود

ساده‌ترین راه نمونه برداری z به این صورت است

$$z = \text{onehot}(\arg\max_i | \pi_i | \pi_i \leq \tau)$$

که $\tau \sim \text{Uniform}(0,1)$ است این تابع نمونه برداری متغیر تصادفی z (max) به طریقی Gumbel را می‌سازد

$$z = \text{onehot}(\arg\max_i \{G_i + \log(\pi_i)\})$$

که $G_i \sim \text{Gumbel}(0,1)$ است

این همان تکنیک reparametrization است حال برای تعریف $\arg\max$ از softmax استفاده می‌کنیم

$$y_i = \frac{\exp(G_i + \log(\pi_i)/\tau)}{\sum_j \exp(G_j + \log(\pi_j)/\tau)}$$

τ به اندازه‌ای است که متغیر تصادفی z به یک احتمال نزدیک می‌شود

$$\tau \rightarrow 0 \Rightarrow \text{softmax} \Rightarrow \arg\max$$

۲) (ج) اگر چه مستند نه معانی نهان: z قطار یک فاکتور توکیه گفته حساس باشد و سبب فاکتورهای دیگر غیر حساس باشد
توسیم بازمانده disentangled است مابین این است که این بازمانده تفسیر نهایی سببی بزرگی و کمای مختلف دارد.

VAE-B یک تغییر VAE با ناگه خاموشی است. فضای disentangled همان است. مشابه فاصله VAE به دنبال max کردن احتمال تولید داده واقعی هستیم (در فضای نهی خالص فاصله بین posterior تجزیه و ترکیبی درونی نهی باشد) پس داریم

$$\max_{x \sim D} \mathbb{E}_{z \sim q_{\theta}(z|x)} [\log P_{\theta}(x|z)]$$

حال ب³ طایفه ای صحراهای کربین حبیب دایم داریم که می‌فرماید

subject to $D_{KL}(q_{\theta_0}(z|x) || p_0(z)) < \epsilon$

با فرض kT بدست لاکر استین $\ln \frac{1}{G}$ می رسد ایم

$$F(\theta, \phi, \beta) = E_{z \sim q_\theta(z|x)} \log p_\phi(x|z) - \beta (D_{KL}(q_\theta(z|x) || p_\theta(z)) - \delta)$$

$$= E_{z \sim q_\theta(z|x)} \log p_\theta(x|z) - \beta D_{KL}(q_\theta(z|x) || p_\theta(z)) + \beta \delta$$

$$= E_{z \sim q_{\theta}(z|x)} \log p_{\theta}(z|z) - \beta D_{KL}(q_{\theta}(z|x) || p_{\theta}(z)) + \beta \mathbb{E}_{z \sim q_{\theta}(z|x)} \log p_{\theta}(z|x)$$

$$\geq E_{x \sim q_{\theta}(z|x)} \log p_{\theta}(x|z) - \text{BD}_{KL}(q_{\theta}(z|x) \| p_{\theta}(z))$$

این تابع خای VAR-B، شکل زیر می آید (البته توجه داریم که B ، ϕ و θ هر دو بستار هستند)

$$L_{\text{Beta}}(\phi, \beta) = -E_{z \sim q_\phi(z|x)} \log p_\theta(x|z) + \text{BD}_{KL}(q_\phi(z|x) || p_\theta(z))$$

وقتی $\beta = 1$ است همان VAE است. دایره β باشد یک شرط محدود کننده روی فضای پتانسیل می باشد. بازبینی z داریم.

بین اعصاب B فزونی با ۲ صافی پسته درک طرح و *disentaglene* شک دارم .
پنهان

فهرست هیچ وقت نمی توان گفت یک مدل GAN یا VAE بهتر است بلکه بسته به کاربرد متفاوت است.

GAN, VAE بر اساس توابع loss مختلفی یاد می گیرند و کاربردهای مختلفی دارند.

یک شکل VAE این است که بعضا تصاویر تار یا blurry تولید می کند اما GAN همچنین مشکل دارد. یک شکل GAN این است که در convergence می شود و متوقف می شود (همان ابهامی که بر اساس آن VAE شکل گرفته) پس هیچ وقت یک راه حل نیست. آل برای تمام مسائل نداریم اگر هم تولید می کند ولی کاربر است به سلیقه VAE می رسم اگر هدف تولید تصویرهای واقعی است به سلیقه GAN می رسم.

$$\text{Relu}(x) = \begin{cases} x & x > 0 \\ 0 & x < 0 \end{cases} \quad \text{Relu}'(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

$$L\text{Relu}(x) = \begin{cases} x & x > 0 \\ \alpha x & x < 0 \end{cases} \quad L\text{Relu}'(x) = \begin{cases} 1 & x > 0 \\ \alpha & x \leq 0 \end{cases}$$

نمونه مشخص است Relu در شبکه های عمیق ارزیابی می شود که باعث شود backprop نامتوازن باشد و آموزش شبکه به شدت کند شود (به همین شکل dead neuron می گویند).

در حالی که LRelu معین مشکل ندارد و متوازن آن هیچگاه صفر نمی شود پس آموزش شبکه بهتر است.



(۴) الف) mode collapse (که می‌بایست کل توزیع را یاد بگیرد) صرفاً بخشی از توزیع را یاد گرفته است

فردا صرفاً یاد به تولید خروجی‌های محدودی نیست علت این امر آنکه این است که داده درسی تنوع خوبی نداشته و یا کلاس‌ها به شدت unbalanced باشند. بابتی شود مدل اهمیت زیادی به بعضی کلاس‌ها ندهد و صرفاً چند کلاس محدود تولید کند.

راه حل کلی این است که اطلاعات W-GAN است که در سمت دیگران شده و همان جا توزیع می‌دهم

دوم packing: به شبکه discriminator چند درسی می‌دهیم (این ایده در PacGAN معرفی شده)

در آن شبکه چن معتقد بودند که علت mode collapse این است که خروجی شبکه صرفاً به یک درسی وابسته است

واله چند درسی می‌دهیم را مجبور باشد یاد بگیرد تنوع بیشتری را یاد بگیرد

(۴) ب)

جنبه (باز هم) در حال پیدا کردن روشی بهتر برای به دست آوردن توزیع دیکای و درسی هستیم. این جا از یک نسخه استفاده می‌کنیم

که طاقی یا جعلی بودن را تشخیص دهد. در واقع این روش تشخیص فاصله EM را به جای Jensen-Shannon divergence درگاه اصلی می‌دهد

min که بابت آن روشی به نام از مدل اصلی و همچنین گفته شدن mode collapse می‌شود. علاوه بر این

خیم‌های زیادی را خروجی می‌دهد و برای می‌چاک کردن باز است. هاشمیه خواهد بود

۴) ب) همچنین این حالتین دانه این ~~کلیه~~ حفا (T-wasserstein) یک معیار احتمالی اساسی است که مترطامی

صفی داری (برای مثال 7-lip schitz) و دارد، باید این می تواند با توجه به discriminant های ده خوش رضا حسنه

مہینہ شود (خوش مقام بہ این معانہ خرمی آن با تفسیر مدرس ، تفسیرات چندیانی نمی کند معنی خود را مع

(Carl lipschitz

(الف) دو نوع رابطه Pix2pix یک مدل image translation است این شبکه برای ترجمه تصاویر از upair استاندارد شود مثلاً به بی هواشناسی

X و Y شایسته است $F: Y \rightarrow X$ and $G: X \rightarrow Y$ را بدینگونه ای که در اینجا ایده اصلی این است که

این شهر را مثال می‌کنیم که یک مجموعه‌های به هم وصل می‌کند و هر دو به هم نفوذ می‌کنند. bijection به این معنی است.

هماری این سیکل یک pi cycle constant دارد. این است به تصویر خدجی سوله اول به عنوان تصویر مدعی

هذه السماء سود، و قد صبي مدعى بها ما بعد اعلان نزيه به اولى بائنه (اين با استقامه از حقاير cycle

$F(G(x)) = x$ and $G(F(y)) = y$

$$F(G(x)) = x \text{ and } G(F(y)) = y$$

تفاوت با GAN: می‌تواند به دنبال paired باشد

یعنی بہ تمام ویر قبل و بعد از تبدل بدلی کہ در شش میل خفاہ نسبت بہ جای این میل می گردانہ با استخوان

مضبوط تصاویر در صورتی که استخراج از اسلاید و استفاده از اسلاید کی تصاویر تبدیل را اینفو

⑤ ب) به تعلم دوره دارد

① Augmentation: آموزش های با افزایش داده استفاده کنیم به دست سوز های کمتر
دسته های می توانیم آموزش های سخت تر را استفاده کنیم
crop, mosaic و استفاده کنیم

② مقادیر خود را را به یک GAN دیتا بدیم تا مقادیر تصویر تولید کند
نتیجه گیری را به cycle GAN بدیم