

سند ۱: ۱.۱) می خواهیم نشان دهیم که صورت ستد PCA را می توان معادل با کمینه کردن reconstruction error در نظر گرفت. این طبقه در PCA به دنبال بینشیدن کردن

واریانس هستیم. و واریانس capture شده توسط هر بردار وزیر (Principal component)

متاثر است با مقدار وزیر آن. از این میانم PCA \Rightarrow reconstruction error

برابر است با مقدار وزیر آن. از این میانم استفاده شده است.

در نتیجه افزایش واریانس capture شده با PCA دفعاً معادل می شود با داشتن

reconstruction error. در نتیجه به جای آن واریانس را افزایش دهیم، معنی می‌کنیم

$$\min_w \frac{1}{2} \|x - w^T w x\|^2 \quad \text{reconstruction error}$$

s.t. $w w^T = I$

د. آن ماتریسی است که بردارهای وزیر های ماتریس دواریانس دارند

چرا که ماتریس w کوواریانس Data نماینده است، همان نظریه است که از PCA باقی مانده است

حال این objective دیفیقاً معادل loss function (با توجه به التهییش) است

$$\rightarrow \text{loss AE: } \min_w \|x - \hat{x}\|^2 = \min_w \|x - w^T w x\|^2 \quad \text{نحوی داشته ایم} \quad \hat{x} = w^T w x$$

نهایتاً تغایر دارد که این است که نیزیج constraint ای در AE وجود ندارد که در تغایر دارد (معادل نرمال نسبت)

در نتیجه AE نزدیک پایه ای است PCA بیان نموده ایم که این پایه ای است بیان نموده

همان فناوری را span می نماید PCA هم می نماید

۱۴-۲) AE همچنین نویسندگی کرد که لیست بودت می‌آید نامم، در نتیجه AF نوک تردد

بگاند که توربینی داشت لیست تشکیل داده است، تهاجمی می‌گذرد reconstruction loss

ناممیستند. در نتیجه توزیع لیست بود می‌شوند پیچیده باشند. در نتیجه نقاط لیست

تعزیزی بود که همچنین خواهد بود و بسیار نامنوار است در عین پیش شود باشد.

در حالت اول، آنکه ۲ عایق که در train داشتم را با یک ابرنور پیش‌نماییم، نقاط

خارج از ابرنور نمی‌توانند ایلوک مارسیج را تشکیل دهند. جا که در train داشتم دست

همچنین نقطه‌ای برای شبکه داشتم نبوده است و انتقادی هم خوبی آن کامل

غیرقابل پیش‌نمایی باشد

آنکه نقطه دادن ابرنور باشند هم باز ممکن است خروجی decoder از الکترونیک مارسیج ابرنور نمایند

پس از دفعه ابرنور نیز همچنان دلیل ذکر نشده توجه AE به سرعت پیوسته میان نقاط داشت

ابرنور پیش نشده است دستگاه است به معنی توجه بسیاری از بخش دیگر شود و در نتیجه

خرده محاذاری تولید شود

(۱-۳۳) در این بخش نیاز داریم که $P_{\theta}(x) = \int P_{\theta}(z) p(x|z)$ را محاسبه کنیم. محاسبه این

با توجه به آنکه باید روی ابعاد ۲ ترفتار شود، قابل محاسبه نیست و عالی است

پس منظمه محاسباتی وجود دارد (معنی بادستن $p_{\theta}(z)$ ، $p_{\theta}(x|z)$ ، هر کدام تفصیلی

مورد تقدیر انتقال بسیار زیاد است). همین مطالقه منظمه در مورد inference نامم حل نمی‌شود

یعنی که با لیست یک نمایه را بست آوریم. یعنی $p(z|x)$. طبق فاعل دیگر:

که در نتیجه $p(x)$ ظاهر می‌شود که فرمولیم شامل محاسبه لیست

۱۰۴) در روش mont-carlo تخمین آنکارال $p(x)$ نامناسب است زیرا مقدار $\int p(x) dx$ ممکن است بزرگ شود.

تخفیف نمایم. برای این کار استاد آنکارال نوشت: شرطی صورت بازگشایی تخمین:

$$p(x) = \int p(z) p(x|z) dz \stackrel{\text{LOTUS}}{=} E_{z \sim p(z)} [p(x|z)]$$

حال از آغاز دانهای این دستم برای آنکارال اید را تخفیف نمایم، تجربه خوب این است که، همراه های در اختیار را میانلئیں نلایم. برای این کار از توزیع $P(z)$ که توزیع نرمال

در نظر گرفته شود میتوان نمونه گرفت و سپس با $p(x|z)$ را محاسبه کرد.

$\frac{1}{N} \sum_{i=1}^N p(x_i|z_i)$ با تقریب N نمونه، میتوان اید را بدین ترتیب تخمین زد:

$$z_i \sim N(0, I)$$

حال با تخمین $p(x|z)$ ، میتوان با روش maximum likelihood به جمعیت بودن پرداخت و θ مطابق را پیدا کرد.

مشکل ایشان که بود داد این است که اگر تعداد نمونه های کافی نباشد، تخمین سیار نویزی خواهد بود. برای تخمین تعداد نمونه های کافی نیز، می تواند پیچیدگی ریاضی بالای داشته باشد

در واقع تعداد نمونه های ریاضی با این علاوه لیست دارد.

$$z : \text{بعد } d$$

۱۰۵) مشکل ایشان که در صورت قبل داشتیم، مشکل در نمونه گیری وجود آوردن z ها با همایش فرازدیده بود. حال بنظر می آید اگر بتوانیم $p(z|x)$ را داشته باشیم، میتوان بمنظور نمونه های وردی

روش monte-carlo از z های که از توزیع $p(z|x)$ برآمدند می آید اتفاقاً کمی

بعضی x هایی که داریم را بگذاریم و z هایی که از آنها می داشیم داشتم. برای نمونه تقریب θ که توزیع ابتدا لازم داریم PDF و CDF آنرا داشته باشیم. دل مبنی نیز داریم

که یعنی در اینجا بـ خود $P(x)$ دنبـاره نیاز
میـدا مـشود کـه اعـدـا به دـبـارـه این
عبـارت رـا محـاسبـه نـیـم و مـقـدـار آنرا زـدـارـم.

در نـیـتـه با عـیرـقـابـل محـاسبـه بـورـن CDF، PDF
نمـوـنـه تـرـفـن اـزـکـان تـوزـع نـیـتـه قـابـل اـنـیـم
نـیـتـه

همـچـنـی در الـلـوـرـیـتم EM، در مرـد E نـیـاز دـرـیـم $P(z|x, \theta)$ رـا محـاسبـه نـیـم کـه هـمانـظـرـه کـه

ایـن عـبارـت قـابـل محـاسبـه نـیـتـه و در تـهـجـین زـدن آـن دـمـبـه سـهـلـشـنـه خـورـدـیـم

$$\log P_\theta(x) = E_{z \sim q_\phi(z|x)} [\log P_\theta(z)] \quad (1.6)$$

$$\begin{aligned} &= \int_z q_\phi(z|x) \log P_\theta(z) dz = \int_z q_\phi(z|x) \log \frac{P_\theta(x,z)}{P_\theta(z|x)} dz \\ &= \int_z q_\phi(z|x) \log \frac{P_\theta(x,z)}{q_\phi(z|x)} - \frac{q_\phi(z|x)}{P_\theta(z|x)} = \int_z q_\phi(z|x) \log \frac{P_\theta(x,z)}{q_\phi(z|x)} + \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{P_\theta(z|x)} \\ &= \int_z q_\phi(z|x) \log \frac{P_\theta(x|z) P_\theta(z)}{q_\phi(z|x)} dz + KL(q_\phi(z|x) || P_\theta(z|x)) \end{aligned}$$

$$\begin{aligned} &= \int_z q_\phi(z|x) \log P_\theta(x|z) dz + \int_z q_\phi(z|x) \log \frac{P_\theta(z)}{q_\phi(z|x)} dz + KL(q_\phi(z|x) || P_\theta(z|x)) \\ &= E_{z \sim q_\phi(z|x)} [\log P_\theta(x|z)] - KL(q_\phi(z|x) || P_\theta(z)) + KL(q_\phi(z|x) || P_\theta(z|x)) \end{aligned}$$

به با توجه به آنکه KL هـمـراـه ثـبـتـه است در نـتـیـجه مـدـپـاسـی برـای عـبارـت بالـا

عبـارت است $E_{z \sim q_\phi(z|x)} [\log P_\theta(x|z)] - KL(q_\phi(z|x) || P_\theta(z))$

عبـارت KL آخر جـون $P_\theta(z|x)$ دـارـه ظـاهـرـه نـسـه دـوـد غـيرـقـابـل محـاسبـه بـورـن

دـدر پـیـاسـی آـنرا بـطـیـه آـن لـیـسـیـضـه مـیـلـیـم

$$KL(P||Q) = E_{x \sim P} \left[\log \frac{P(x)}{Q(x)} \right] = \int \log \frac{P(x)}{Q(x)} P(x) dx = 1 - V$$

$$\rightarrow KL(N(\mu, \Sigma) || N(0, I_d)) = \int \frac{1}{2} \log \frac{1}{|\Sigma|} + \frac{1}{2} \underbrace{(x - \mu)^T \Sigma^{-1} (x - \mu)}_{N(\mu, \Sigma)} \times p(x) dx$$

scalar $\rightarrow \text{tr}(\text{Scalar}) = \text{scalar}$
 $\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$

$$= \frac{1}{2} \log \frac{1}{|\Sigma|} \underbrace{\int p(x) dx}_{1} + \frac{1}{2} E \left[-\text{tr}(\Sigma^{-1} (x - \mu)(x - \mu)^T) + \text{tr}(I^{-1} x x^T) \right]$$

$$= \frac{1}{2} \log \frac{1}{|\Sigma|} + \frac{1}{2} E \left[-\underbrace{\text{tr}(\Sigma^{-1} \Sigma)}_I + \text{tr}(x x^T) \right]$$

$$= \frac{1}{2} \log \frac{1}{|\Sigma|} - \frac{n}{2} + \frac{1}{2} E \left[\underbrace{\text{tr}(x x^T)}_n \right] = \text{tr}(E_{N(\mu, \Sigma)} [x x^T])$$

$$= \frac{1}{2} \log \frac{1}{|\Sigma|} - \frac{n}{2} + \frac{1}{2} \text{tr}(\Sigma + \mu \mu^T)$$

$$= \frac{1}{2} \log \frac{1}{|\Sigma|} - \frac{n}{2} + \frac{1}{2} \text{tr} \Sigma + \frac{1}{2} \text{tr} \mu \mu^T$$

با توجه به آنکه Σ متفاوت ایست، عناصر تغیری آنرا به درجه تغییری بینم. و عناصر متفاوت

$$\frac{1}{2} \left[\sum_{i=1}^d -\log \sigma_i^2 - 1 + \sigma_i^2 + \mu_i^2 \right] \rightarrow \text{دترمینان ماتریس متفاوت: حاصل قدرت عناصر متفاوت} \rightarrow N_i \rightarrow \mu_i \rightarrow \mu \mu^T$$

دترمینان ماتریس متفاوت: حاصل قدرت عناصر متفاوت

$$\Rightarrow p(x|z) = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{1}{2} (x - \mu)^T (x - \mu)} \quad p(x|z) \sim N(\mu, I) \quad (1.8)$$

$$\Rightarrow E_{z \sim q(z|x)} [\log p(x|z)] \xrightarrow{\text{یادآوری}} \log p(x|z) = \log \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{1}{2} \|x - \mu\|^2}$$

$$= \frac{1}{(\sqrt{2\pi})^d} \log e^{-\frac{1}{2} \|x - \mu\|^2} = \frac{1}{(\sqrt{2\pi})^d} \cdot -\frac{1}{2} \underbrace{\|x - \mu\|^2}_{\text{MSE}}$$

$$p(n|z) \rightarrow \text{Bernoulli} \rightarrow p(n|z) = \prod_{i=1}^k p_i^{x_i} (1-p_i)^{1-x_i}$$

$$\log p(x|z) = \sum_i x_i \log p_i + (1-x_i) \log (1-p_i)$$

است x one-hot $\in \mathbb{R}^k$

Encoder \rightarrow (1.9) توزیع $p(z|x)$ را منظمه بادلیم. و سپس از آن نمونه بگیریم و به Decoder

بردهیم. آنرا خود توزیع $p(z|\mu, \Sigma)$ نموده بگیریم که $N(\mu, \Sigma)$ است، هسته

نمی توانیم از واحد نمونه نیز ترددان را مبین دهیم. backprop

دایی آنده اند و خود (μ, Σ) سهیل بگیریم، از نرمال استانداره نموده می گیریم و سپس آنرا

$$z = \mu + \sigma \cdot \text{sample}$$

مقدار مورد نظر اسقل می دهیم: بدین صورت:

حال در backprop، با داشتن ترددان z ، می توان ترددان را داشت μ, Σ را می سبک کرد

و تنها کافی است که مقدار خود sample را داشته باشیم تا ترددان را داشت μ, Σ و z

را محاسبه نیم و از آنجا به شکل backprop & encoder (دیگر فستق نسبت به sample) داریم). خود می سبک encoder، μ, Σ قرار می دهیم

سالهای ۲۰۱۸ نسله logistic regression یک binary classification است که اینجا می‌بندد. آنچه خوبی الگوریتم آنرا محدود نماینده توزیع احتمال است که D دلخواه بگیریم، تابع هزینه «آن بین صورت است:

$$-\sum_{i=1}^N \log(D(x^{(i)})) - \sum_{i=1}^N (1-y^{(i)}) \log(1-D(x^{(i)}))$$

که در آن $(y^{(i)}, x^{(i)})$ داده‌نامه دیتابست است. همچنین با توجه به این

نوع سند binary است، هر داره تنها یک از عبارت‌های $y^{(i)}$ و $1-y^{(i)}$

یک خواهد بود و دیگری صفر خواهد بود. هنین به معنای هزینه یا $\log D(x^{(i)})$ محاسبه خواهد شد

اگر "صفر" بودن آنرا مستغص می‌لند. پس در logistic regression آن $y^{(i)}$ و سعی می‌شود

عبارت $D(x^{(i)})$ مالسیم شود، یعنی احتمال بیشتر مالسیم شود و آن $1-y^{(i)}$ ، احتمال ایک

شیک نماید می‌شود (احتمال صفر شون بینیش شود)

$$z^{(i)} \rightarrow P(z) \quad (2) \quad x^{(i)} \rightarrow P_{\text{Data}} \quad (1) \quad \text{حال در GAN درودی دوختن دارد:}$$

و عبارت loss پین صورت است:

$$+ E_{z \sim p(z)} [\log(1 - D_{\theta_D}(G_{\theta_G}(z)))]$$

الر عبارت اید ریاضی را با بیانی تجربی فرم، عبارت loss بسیار عبارت logistic با همان

cross entropy نماید. در اینجا هم ب دنبال مالسیم بودن احتمال واقعی بودن (x) .

همین روشی که داره واقعی باشد) و ب دنبال می‌بینیم بودن احتمال واقعی بودن هستیم (در صورت که

داده تولید شده باشد). در logistic هم آن $y^{(i)}$ را نهاینده واقعی بودن نمایه دلخواه بگیریم

دنبال نمی‌کرد هستیم. نهایندر GAN دو نشانه وردی وجود دارد (برخلاف logistic

دنبال نمی‌کرد هستیم. نهایندر GAN دو نشانه وردی وجود دارد (برخلاف logistic

دنبال نمی‌کرد هستیم. نهایندر GAN دو نشانه وردی وجود دارد (برخلاف logistic

پس در واقع D یک دسته بند است که همراهان با G در حال train می‌باشد است

۲۰۲: G دنبالهای تابع $\log(1-D(G(z)))$ است. در این تابع D مقدار نزدیک به ۱ است یعنی در این سیار تردید و نزدیک به صفر است.

نقطه نزدیک به صفر در این تابع در واقع نهایانه تقاض است که G بدلیل کرده است

و در نتیجه D آنرا شخص دارد است. در این حالت G بیاز دارد که در اینها و اپراتور خوب بسته اور تا بتواند خود را اصلاح کند که در این نقطه بسیار صعب است.

اما $\sum_{i=1}^N \log(D(G(z_i)))$ ، مستقیماً تابع $\log \sum_{i=1}^N D(G(z_i))$ در نقطه نزدیک به صفر، بزرگ است

و در نتیجه G این فرمت را من بعد تا خود را در طول train بخود بخواهد (در حالی که در 10^{55} اول تازه در بایی G مفونه های خوب تولید من لند در این بزرگ می شود)

که بدد نمی خورد

۲۰۳: در اینجا ابتدا G را ثابت می نیزیم و نسبت به D بهینه سازی انجام می دهیم (در واقع این موضع

$$V(D, G) = E_{x \sim P_{\text{Data}}} [\log D(x)] + E_{z \sim P_z} [\log (1 - D(G(z)))]$$

از راهیت بازی ناشی می شود) :

$$\Rightarrow V(D, G) = E_{x \sim P_{\text{Data}}} [\log D(x)] + E_{x \sim P_g} [\log (1 - D(x))] \\ = \int_x P_{\text{Data}}(x) \log D(x) dx + \int_x P_g(x) \log (1 - D(x)) dx \\ = \int_x (P_{\text{Data}}(x) \log D(x) + P_g(x) \log (1 - D(x))) dx$$

$$\Rightarrow D^* = \arg \max_D \int_x (P_{\text{Data}}(x) \log D(x) + P_g(x) \log (1 - D(x))) dx$$

حل با توجه بر آنکه D محدودیت ندارد، آنرا ب تغیر مجهول می توان در نظر گرفت.

حل تابع $y \cdot \log y + b \cdot \log(1-y)$ به صورت زیر بهینه شود:

$$\frac{\partial}{\partial y} a \log y + b \log(1-y) = \frac{a}{y} - \frac{b}{1-y} = 0 \Rightarrow (a+b)y = a \Rightarrow y = \frac{a}{a+b}$$

$$\Rightarrow D^* = \frac{P_{\text{data}}}{P_{\text{data}} + P_g}$$

حال فرض می کنیم استراتژی D ناکت است (نهان استراتژی بهینه اش است) و بهینه سازی را برای G

$$G C(G) = \max_D V(D, G)$$

$$\begin{aligned} & \max_D \int_x (P_{\text{data}}(x) \log D(x) + P_g(x) \log (1-D(x))) dx \\ &= \int_x \left(P_{\text{data}}(x) \log \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_g(x)} + P_g(x) \log \frac{P_g(x)}{P_{\text{data}}(x) + P_g(x)} \right) dx \\ &= \int_x \left(P_{\text{data}}(x) \log \frac{P_{\text{data}}(x)}{2} + P_g(x) \log \frac{P_g(x)}{2} \right) dx - \log 4 \\ &= KL(P_{\text{data}}(x) || \frac{P_{\text{data}}(x) + P_g(x)}{2}) + KL(P_g(x) || \frac{P_{\text{data}}(x) + P_g(x)}{2}) dx - \log 4 \\ &= 2 JS(P_{\text{data}}(x) || P_g(x)) - \log 4 \end{aligned}$$

پس دلیل معیار KL در نسبت JS ، همراه بزرگتر ساوی صفر هستند همچنین مینیمم عبارت بالا

برای $\log 4$ - خواهد بود یعنی زمان که JS صفر شود، این زمان رخ من درد که در توزیع

$$P_g = P_{\text{data}}$$

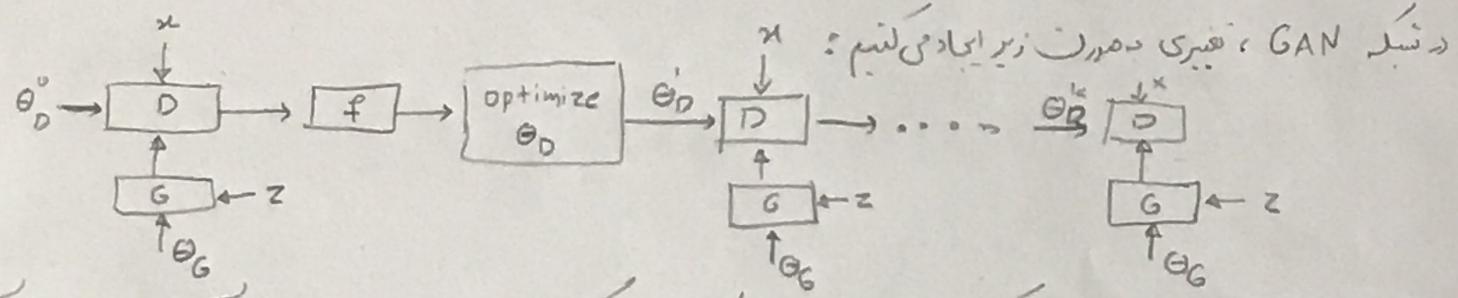
سند ۲۰۴ : این مدل عبارت است از وقتی که نوزیع جزوی GAN می‌باشد صریح باید داشت
دارای چند قله یا mode باشد، اما GAN تنها بین از مدل‌های آنرا باید درست نماید، در نتیجه
وقتی نمونه‌های مختلف از نفسی لیست ترتیب می‌شود، نمونه تولیدی حاصل از آنها بسیار شبیه و میزان
است، چرا که GAN نتوانست هست آنترپیک نوزیع معقصد را به خوبی مدل کند و نمونه‌های
حاصل، همه از یک mode باشند. حال بررسی می‌شود که چرا میان انفاق در چهار چوب یا چهار GAN
یک نوزیع معقصد دارد که آنرا در نظر می‌گیریم که در آندازه D_{train} ، نمونه‌های تولیدی به کلی اول نوزیع می‌توانند
تولید شوند. در این شرایط D ، ب نقاط تولید قله « D » انتبازگو بالای می‌دهد و چرا که تنها نمونه‌های داعی
را از آنها دیده است. در مرور آن اول چون ترکیب از نمونه‌های داعی و ساختنی وجود داشته است که سمعت نرخی
دشیز دارد و خامدنا $D(x) > D_{\text{train}}$. حال در این شرایط در مرحله تعلیم، چنان که از استراتژی D
با خواست، تمام نمونه‌های تولیدی خود را به سمت تله D می‌برد تا سود خود را مال‌سازی کند. لیکن
نارخ از لیسته درودی، معنی بی‌لذت بی‌رفتگی خاص را آنها مدل کند. و زبان در این حالت
است لزومی به تغییر آیدست پارامترهای خود نمی‌بینید چرا که دادا آن نم است. D در اینجا دارای سلسله
بری شناسایی این انفع خواهد داشت و D زیاد را می‌بیند می‌شود. بعد از آنکه D پارامترهای خود را
آپدیت کند و تواند تضییع دهد که توسط G کل خوده است. اما حال G دوباره منزد باشان
استراتژی می‌شود D کل نزد و حالا فاعل از حسنه قله اول تولیدکننده (چون در حال حاضر قله یک تولیدکننده
 D نقاط داعی نسبت می‌فرمود). در واقع این امر باعث می‌شود که داده لذت زمان در training
به جای آنکه G کل مدل‌ها را باید بگیرد، از مدل‌ای به تله دیگر نیز دارد و به اصلاح بازی موش

و تربیت میان G و D رخ دهد

برای حل این مشکل راه حل‌های تفاوتی بیان شده است. اینها دو صور را به انتشار توضیح می‌دهند

روش اول: minibatch discrimination : درین "ش" D را که آنده تها با دیگر یک سهل ورودی دسته‌بندی خود را ایام دهد، یک minibatch از نمونه‌ها که اختیار داشته باشد و دسته‌بندی را بر این اساس انجام دهد. یعنی یک معیار شایسته میان یک نمونه و نمونه‌ها که دلیر در آن minibatch تعریف شود، حال اگر صیل یک هدف نمونه و نمونه‌ها دلیر در minibatch شایسته غیرعادی اساس نکند، بفهمد که حالت خارج از راکوردن بزرگ، و در نتیجه این امر دشمنانشان نکند. این طریق را به این سهیت منبرد که نمونه‌هایی که تولید می‌کنند از تنوع مناسبی بخوردار باشند و سُلول mode collapse را تلاع می‌کنند.

روش دوم: unrolled GAN : در صورت اینکه mode collapse در یک mode بهینه پیدا می‌کند و تا برداشی که D این مسند را لشکر نکند، استراتژی (پالامتر) خود را تغییر نمی‌دهد، چون عدها آن تهیبا براساس استراتژی فعلی D و محاسبه می‌شود. حال اینه این است که اگر در یک بازی بازمان از استراتژی‌های آینده ریسیب (با تابوت بودن استراتژی خوش) آن باشد، استراتژی بهتری انتخاب می‌کند.



در واقع استراتژی G را تابوت می‌کنیم و D را بار آپدیت می‌کنیم تا استراتژی‌های او در آینده را برسیت آیم. در آن بحدله ایم، D یکبار آپدیت می‌شود (برنسب مردمه کام)، اما G براساس chain rule بهترین تهیم که مرد یکجا آپدیت می‌کند (متابه rnn). این بدان معناست که G در آپدیت خود تمام این استراتژی D در در آینده را بدقت قرار داده است تا فهمیم بهتری بگیرد. حال سُلول mode collapse (mode) بین صورت حل می‌شود که G می‌بیند با استراتژی فعلی تولید نمونه از یک

D در مراحل بعدی این روش را مستحب می‌نمود و سود او را افزایش خواهد کرد. در نتیجه این استراتژی را انتخاب نمی‌کند و به سهیت تولید نمونه‌های تنوع می‌کند.

۲.۵) الف) قبل از تعریف فاصله بین دارایم با عکس روشن جایگان را
 آشنا شویم. روشن جایگان (y, x) که به مدد کننده بیندر میزان جرم (mass) جایگاه بین
 دو نقطه y, x در فضای است. مصادق های مختلف و تاریخ برای نظر
 وجود دارد. به عنوان مثال یک ایش در تعدادی از نقاط تعدادی سرباز دارد و حالا می خواهد آنها
 ۳ در نقاط جدید مستقر کند. هر mapping ای که این کار را ممکن می کند
 است. واضح است که هر plan هریک خاص خود را خواهد داشت. هریکی plan خاص
 عبارت است از :

$$\text{plan} \quad T(c, \gamma) = \int_{\text{هسته کل آن}} c(x, y) \gamma(x, y) dx dy = E_{\substack{\text{آنکه} \\ \text{آنکه}}}[c(x, y)]$$

$\gamma(x, y)$ هریک انتقال میان y, x
 داره شده میان y, x

حال فاصله Wasserstein دری دو توزیع را منظمه میسیم. توزیع تناصر با پنجه سربازها
 Wasserstein P_G داری دو توزیع آنرا زید. حال برای دو توزیع P_G و P_{Data} فاصله
 دو نقطه است که مثلا آنرا زید.

$$W(P_G, P_{Data}) = \inf_{\gamma \in \Pi(P_G, P_{Data})} T(c, \gamma) \quad \text{عبارت است از :}$$

د) $T(P_G, P_{Data})$ عبارت است از خانواده $c(x, y)$ میان های مختلف میان این دو توزیع. در واقع این
 عبارت دنال فاصله ای است که هریک انتقال میان دو توزیع را کمینه نماید. ولی اگر (y, x) را فاصله انتقالیس

$$W(P_G, P_{Data}) = \inf_{\gamma} E_{\substack{\text{آنکه} \\ \text{آنکه}}} [\|x - y\|_2] \quad \text{در نظر بگیریم داریم :}$$

$$1) \int \gamma(x, y) dx = P_{Data} \quad \text{عبارتند از :} \quad \int$$

$$2) \int \gamma(x, y) dy = P_G$$

حال نماید این محدود است، میتوان آنرا به ششم یک مسئلہ Primal نوشود.

$$\text{آن مسئلہ Primal} \rightarrow \text{ملت} \int_{x, y} \gamma(x, y) \text{ قابل محسنه است. حال با نوشت مسئلہ Dual}$$

$$W(P_{Data}, P_G) = \sup_{x \sim P_{Data}} E[f(x)] - E[f(x)]_{P_G}$$

دین \tilde{f} از تغیرهای مبتدا dual objective ظاهر می شود.

پذیره ای داشت f تابع

$$\frac{f(x_1) - f(x_2)}{x_1 - x_2}$$

در طالث پذیری

$$E[\log(1 - D(G(z_i)))]$$

مشتل gradient vanishing ایجاد می شود

در واقع، بیشتر بود، بازی برای نسق تابع صحفی می شود.

مشکل آن ترجیح دارم. اما این تابع نیز مشکل دارد.

اما در WGAN ها با توجه به آنکه تابع هوف f یک تابع ایجاد می شود.

اولاً طریق آن طبق می باشد و همچنین در عمل دیده می شود که در قالب،

چه خوب کار نمی کند، در این متناسبی دارد و در نتیجه در train GAN، فسللاتی

از بیلability training یا حساس بودن برآفخ می شود

(۱) KL، تفاوت نیست که حالت و قیمت به دنبال یعنی درون دو توزیع P و Q دستیم، می فرضیم

فاصله دو توزیع صفر سرد که KL این رفتار را ندارد. موضوع دیگر این است که KL رسید

نقطه مشترک دو توزیع در دری آن تعیین می سرد. حال که ابعاد بسیار بالا، اضطراب آن نقطه مشترک

دو توزیع تهی باشد بسیار بالاست و در نتیجه KL آنها تعیین نمی شود. این موضع در کل هم وجود دارد

دون آن هم براساس KL است، اما در Wasserstein این اتفاق نمی دهد

موضوع دیگر هم در این ها بهتر است که در ج تولیخ داده شد

مسئلہ ۳۰۱ . الف) - توابیه موسيقی فساد بایک علس : شبلهای تکرار شونده یک بیضیز
چرا که ورودی تنها یک علس است که یک علس بسری زمان نبست ، پس ورودی میں است
اما خوبی یک موسيقی است که موسيقی یک سری زمان است (یک سیلیال صوت در مجموع زمان)
و در نتیجه نیاز داریم هر دوین خوبی در زمان های مختلف از شبله بلیریم تا موسيقی ماسیل دهد

- شخصیعنی با اساسی یک حق : شبلهای تکرار شونده چند بیک : ورودی یک حق است

که با یک سری زمان شناخته است ، چرا که لفمات دیک حق به لفمات خافر در مدل و بعد از آن
لفمات در حق و مدل است . مدد نیخ ورودی شبله را چند تایی دنگر می بلیریم ، که کل زبان ملزم
محل مدل است . چون مسئلہ یک مسئلہ دسته بندی است ، تنها در زمان انتها ، بعد از دیدن کل
جمله ، من زبان تعبیم درست که sentiment جمله چیست . مدد نیخ خوبی شبله میں است

- تعیین نقش لفمات در یک جمله : شبلهای تکرار شونده چند بیک نامهمان

ورودی یک مدل است که شناخته با یک سری زمان است . خوبی نیز با قوام بـ آنکه برای هر لفم

باید نقش آنرا تعیین کنیم چند تایی است . نامهمان از این قوای در نظر در نظر نشده است که نقش حملات

من آواند بـ کل جمله و اساسی باشد نه بـ لفمهای دیده شده نـاـاـان . نـایـدـیـشـتـرـنـ ترـیـشـنـ شـبـلـ (bidirectional)

- ترمیم یک حق : شبلهای تکرار شونده چند بیک نامهمان : ورودی یک مدل است که یک سری زمان
است پس چند تایی است . خوبی هم یک مدل است بـ زبان مقصود که آن هم سری زمان است .

نـاـمـهـمـانـ است چرا که برای ترجمه نیاز است کـلـ جـمـلهـ رـاـ دـیدـهـ باـمـیـمـ

ب) bidirectional : در یک شبکه rnn معمول ، اطلاعاتی که در یک رشته + دلیم حاصل از آنها اطلاعات از زمان بُعد تا اینست ، یعنی زمان هایی که تابعی دیده ایم . اما در لایه برخانی نیاز داریم که از گذشته بعدی نیز اطلاعات داشته باشیم تا بهترین مدلر را داشتیم . جلیکه داده
حالی context می دویم . فعلاً تعیین نقش کلمات در یک جمله نیاز دارد که اطلاعات آینده را
بداند . یعنی کلمات حاضر شده بعد از آنها مرور نظر را بسیند و سپس تفہیم لیری کند . نسل کنیم
است که نیاز داریم که sequence از طبقه را برای پیش‌لشون داشته باشیم و برقرار است online
نهی تولک تفہیم لیری کرد

Deep RNN : آر تابع که در هر لحظه RNN فراست می‌شود پیشیده باشد و همین درون
لایه بازگشتی به یادگار توابع پیشیده توکن می‌نماید و اینها بازگیری آنها را مراهم می‌نمایند.
همچنین می‌توانیم شبکه عصب عمیق، اندازش عمق باعث می‌نماید که ساختارهایی می‌شوند که
ویژگی‌های معنایی آنرا استخراج کنیم. مثلاً در لایه n_{RNN} که ورودی‌های آن حروف الفبا است
در لایه اول من توان حرف صدادر از بنده را تسفیف نمایم. در لایه دوم براساس اطلاعات لایه اول،
من توان تسفیف نمایم که بعد از این حرف صدادر، یک حرف بی صدا من آمده و به همین ترتیب ...

$$f_t = - \sum_{j=1}^c y_{tj} \log z_{tj} \quad \text{ابدا } f_t \text{ را بدين صورت تعريف من نيم:} \\ f = \sum_{t=1}^T f_t \quad \text{نم تعداد دستورها يا ايجاد } z_t \text{ است. حال } f \text{ مى سرد:}$$

حال از این متن $\frac{1}{2}$ را سنت به مقادیر فراسته $\frac{1}{2}$ حاسسه می نماییم و متنق $\frac{1}{2}$ حاصل جمیع می شود.

$$* \rightarrow \frac{\partial L_t}{\partial x_t} \rightarrow dt \text{ تهاونهات } \rightarrow \text{ همان سبق} \\ \text{سنتق وجود ندارد در بقیه موارد} \rightarrow \text{ cross entropy} \\ \text{سنتق } \rightarrow \text{ logit ها} / \text{ صفر است} \\ \text{که به ترتیب قبل آنرا اثبات نمودیم}$$

$$\Rightarrow \frac{\partial \mathcal{L}_t}{\partial \alpha_t} = (z_t - y_t)$$

$$* \rightarrow \frac{\partial \mathcal{L}_t}{\partial W_{hz}} = \frac{\partial \mathcal{L}_t}{\partial \alpha_t} \cdot \underbrace{\frac{\partial \alpha_t}{\partial W_{hz}}}_{\text{مشتق ضرب ماتریسی}} = (z_t - y_t) h_t^T$$

$W_{hz} \in \mathbb{R}^{C \times H}$

$$* \rightarrow \frac{\partial \mathcal{L}_t}{\partial b_z} = \frac{\partial \mathcal{L}_t}{\partial \alpha_t} \cdot \underbrace{\frac{\partial \alpha_t}{\partial b_z}}_I = (z_t - y_t)$$

$b_z \in \mathbb{R}^C$

$$\frac{\partial \mathcal{L}}{\partial W_{hh}} = \sum_t \frac{\partial \mathcal{L}_t}{\partial W_{hh}} \rightarrow \frac{\partial \mathcal{L}_t}{\partial W_{hh}} \stackrel{\text{chain rule}}{=} \frac{\partial \mathcal{L}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial W_{hh}}$$

: پس از تبدیل به آن داده های موردی را که $W_{hh} \cdot h_t$ دار است داریم

$$\frac{\partial h_t}{\partial W_{hh}} = \sum_{k=1}^t \frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial W_{hh}}$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial W_{hh}} = \sum_t \frac{\partial \mathcal{L}_t}{\partial W_{hh}} = \sum_t \frac{\partial \mathcal{L}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial W_{hh}}$$

$$= \sum_{t=1}^T \sum_{k=1}^t \frac{\partial \mathcal{L}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial h_t} \cdot \underbrace{\frac{\partial h_t}{\partial h_k}}_{\text{از مشتق این عبارت بک}} \cdot \underbrace{\frac{\partial h_k}{\partial W_{hh}}}_{\text{این یک عبارت بازگشتنی است}}$$

$$h_t = \tanh(W_{hh} (\dots (W_{hh} h_k + \dots)))$$

پس از این داده های W در مشتق ضرب ماتریسی

$$\frac{\partial h_i}{\partial h_j}$$

ظاهر می شود که نتیجه حاصل از ضرب W

می شود و در نتیجه چون رابطه بازگشتنی است، تعداد زیادی W در نتیجه ضرب می شود

$$\frac{\partial \mathcal{L}}{\partial w_{xh}} = \sum_{t=1}^T \frac{\partial \mathcal{L}_t}{\partial w_{xh}} \rightarrow \frac{\partial \mathcal{L}_t}{\partial w_{xh}} = \frac{\partial \mathcal{L}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial w_{xh}} \quad (2)$$

$$\rightarrow \frac{\partial h_t}{\partial w_{xh}} = \sum_{k=1}^t \frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial w_{xh}}$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial w_{xh}} = \sum_{t=1}^T \frac{\partial \mathcal{L}_t}{\partial w_{xh}} = \sum_{t=1}^T \frac{\partial \mathcal{L}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial h_t} \cdot \sum_{k=1}^t \frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial w_{xh}}$$

$$= \sum_{t=1}^T \sum_{k=1}^t \frac{\partial \mathcal{L}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial w_{xh}}$$

استدلال

مسئله با این ترتیب قابل حل نبوده بود و آنرا در بحث ضرب سعد w_{xh} دیده بیشتر می سود

) می توان نشان داد که اگر ماتریس W به صورت پیش سرمه دید بروار ضرب نمود،

حاصل در نتایج بزرگترین بروار و نزدیک این ماتریس خواهد بود و فریب آن تناظر با λ است

که λ مقدار دینه تناظر با بزرگترین بروار و نزدیک K تعداد انفعانی است که ماتریس ضرب نموده است.

در نتایج دو بخش قبل توضیح داریم که w_{xh} چندین بار در خود ضرب می شود. حال از تفکیر بالا

استفاده می کیم. \rightarrow درایان بسیار \rightarrow فریب بسیار بزرگ $\rightarrow \lambda > 1$
 بزرگ می شود ($\uparrow \lambda$)

$\lambda < 1 \rightarrow$ درایان بسیار \rightarrow فریب بسیار بود $\rightarrow \lambda < 1$
 کوچک می شود ($\downarrow \lambda$)

این نتیجه نسبت به شبکه های عصبی عمیق خاتمه است چنانچه اینجا یک W بیسان دارد در خود

ضرب می شود و تفکیر بالا بزی این حالت بروار است اما در شبکه های عصبی عمیق به عمل W لایه های

مختلف در هم ضرب می شوند و تفکیر بالا دیگر صادر نمیست و از این فریب ضرب های

به ضرب سعد در خود نمیست

(5) در این روش بک ترسولو تبعیض می‌لسم و اجازه نمی‌دهیم اندوه تراویان از این ترسولو
لطفیتر شود. ولی صورت که دهنر مت گامب تراویان، فرم آنرا حساب می‌نمیم و اگر فرم آن
که ترسولو نهیں نمی‌شود تفسیر شود باشد، تراویان را به دلیل normalize \hat{y} لیسم تا از $\text{exp}(-\hat{y})$ نمودن
آن حلولیتر کنیم. اگر ترسولو را به دلیل نظر بگیریم:

$$\text{grad} \leftarrow \text{grad} \cdot \frac{c}{\|\text{grad}\|}$$

اندازه

این کار بادست می‌شود که تراویان دهنر ~~که~~ بیشتر از c نباشد. c های پر پارامتر سُلْم است

(6) پُونده وزنها shared هستند. به عنوان مثال اگر W_{hh} را براسامن زمان نماید

لذیم، تراویان که W_{hh}^T در T می‌لید مناسب و نیز است رجرا دهنر backpropagation شده است) و چون وزنها shared هستند در نتیجه W_{hh} آبلیت مناسب دهمانجا
دیافتگی نمود. اما اگر backpropagation through time تراویان به ندرت می‌کند

این موضع بادست می‌شود که $\text{loss} = \sum_{t=1}^T \text{loss}_t$ از زمان‌های آنرا، بر روی لایه‌های زمان‌های قبل تر
تأثیر نمی‌پذارد. یعنی در داتا نهایا ب حریق زمان‌های اپیزود توجه نمود نه بلطف
زمان. در حقیقت ماقص فراسنیم کل Sequence در نتیجه loss تأثیر پذارند نه فقط

زمان‌های اپیزود. با مشابه تحقیق زمانی از روش‌های مانند truncated backpropagation through time استفاده می‌شود

پلت نهادس و جود ندارد

(۱)

$$\frac{\partial \hat{f}_t}{\partial W} = \underbrace{\frac{\partial \hat{f}_t}{\partial h_t}}_{\text{دراستگا از بازگردان نسبت به } h_t} \cdot \underbrace{\frac{\partial h_t}{\partial W}}_{\text{هر یک کردن چون تابعی دارد}}$$

از دو قسیر نهایی نهاد متن

را نسبت به W داشتلود بینی از h_k روانهای زمانی ای تبیهو یک از c_k روانهای زمانی قبلی

$$\Rightarrow \frac{\partial h_t}{\partial W} = \sum_{k=1}^t \underbrace{\frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial W}}_{\text{عبارت پارسی}} + \underbrace{\frac{\partial h_t}{\partial c_k} \cdot \frac{\partial c_k}{\partial W}}$$

آخر دین عبارت W ظاهر شد، به سیم دخواه
من و سیم

عبارت دیگر ابررسی می شود

دیدیم که این عبارت باعث صرب شدن مقادیر W می شود پس

$$\Rightarrow \frac{\partial h_t}{\partial c_k} = \underbrace{\frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial c_k}}_{\text{جهوت پارسی نسبت به } 0 \text{ بازگردانی خود تابعی}} + \underbrace{\frac{\partial h_t}{\partial c_t} \cdot \frac{\partial c_t}{\partial c_{t-1}} \cdot \frac{\partial c_{t-1}}{\partial c_k}}$$

باشهش

اگر دیگر از ده عبارت ملا W ظاهر نشد دستیه دخواه من و سیم. عبارت «نم» ابررسی می شود:

$$\frac{\partial h_t}{\partial c_t} = o_t (1 - \tanh^2(c_t)) \cdot \frac{\partial c_t}{\partial c_{t-1}} = f_t$$

$$\Rightarrow \frac{\partial h_t}{\partial c_t} \cdot \frac{\partial c_t}{\partial c_{t-1}} \cdot \frac{\partial c_{t-1}}{\partial c_k} = o_t (1 - \tanh^2(c_t)) \cdot f_t \cdot \underbrace{\frac{\partial c_{t-1}}{\partial c_k}}_{W \text{ ظاهر نشد}}$$

عبارت پارسی خامن

لند کردن
دو و آنچه نشان دادیم در LSTM، درستگی نسبت به W ، تبیه از یک راهبردی لند کردن

W در آن هزب نهی شود و ضمیمه کردن RNN وجود داشت با متوجه چیزی