

① ^آمیار FID شباهت بین بازخانی های تصاویری واقعی و تصاویری تولید شده در feature space شبکه inception را اندازه گیری می کند. این بازخانی ها عموماً در سری pooling آخر شبکه inception است. FID نامده Fréchet بین دو توزیع گاوسی چند متغی را اندازه گیری می کند. Fréchet Inception Distance

میار IS یا Inception Score یا به این ایده (است که تصاویری تولیدی با کیفیت مناسب باید احتمال likelihood بالایی تحت مدل های pretrained classifier داشته باشند و همچنین احتمالاً class های مشخصی نیز داشته باشند. برای مقایسه این میار، ابتدا تصاویری تولیدی بر یک مدل و به دست می آید شبکه inception داده می شوند و این دسته بندی توزیع احتمال روی تمامی کلاس ها را میسر می کند. سپس میار IS با استاندارد از میانگین گرفتن exponential entropy احتمالات کلاس ها روی تمامی تصاویری تولیدی مقایسه می شود.

ب) بعضی از محدودیت های این میار عبارتند از: هزینه محاسباتی بالا / نیاز به دیتا بزرگ / عدم تعمیم پذیری مناسب و ... یک مثال از این محدودیت ها زمانی است که می خواهیم تصاویری تولیدی از چهره های انسانی را مورد ارزیابی قرار دهیم. اگر مدلی داشته باشیم که تصاویری با کیفیت بالا تولید می کند اما تنوع بالایی ندارند و مثلاً همه آنها چهره های جوان و با صورت مستطیل هستند، آنگاه FID نمی تواند این عدم تنوع موجود تنوع را مشخص کند و منجر به این مدل به بعد.

ج) این میار به این ایده است که تصاویری تولیدی نه تنها باید شبیه تصاویری واقعی باشند، بلکه باید مدل های مولد بتوانند بازه تنوعی از توزیع داده ها را پیش بینی دهد. برای مقایسه این میار، ابتدا مدل مولد تعداد زیادی تصویر تولید می کند. سپس این تصاویری تولیدی با استاندارد از یک pretrained encoder به یک feature space نگاشت می شوند. میار Coverage با استاندارد از محاسبه convex hull آن فضای نگاشت است. این تصاویری درون feature space به دست می آید. میار Diversity با استاندارد از محاسبه میانگین فاصله بین این تصاویری embedded به دست می آید. استاندارد بالا در این میار به این معنیست که مدل مولد بازه تنوعی از توزیع داده ها را پیش بینی کرده و همچنین به صورت مکرر در feature space قرار دارند.

$$\Omega = E_{x \sim P_{\text{data}}(x)} [E_{q_{\phi}(z|x)} [\ln P_{\lambda}(z) - \ln q_{\phi}(z|x)]] \quad (2)$$

$$= \int P_{\text{data}}(x) \int q_{\phi}(z|x) [\ln P_{\lambda}(z) - \ln q_{\phi}(z|x)] dz dx$$

$$= \iint P_{\text{data}}(x) q_{\phi}(z|x) [\ln P_{\lambda}(z) - \ln q_{\phi}(z|x)] dz dx$$

$$= \iint \frac{1}{N} \sum_{n=1}^N \delta(x - x_n) q_{\phi}(z|x) [\ln P_{\lambda}(z) - \ln q_{\phi}(z|x)] dz dx$$

$$= \int \frac{1}{N} \sum_{n=1}^N q_{\phi}(z|x_n) [\ln P_{\lambda}(z) - \ln q_{\phi}(z|x_n)] dz dx$$

$$= \int \frac{1}{N} \sum_{n=1}^N q_{\phi}(z|x_n) \ln P_{\lambda}(z) dz - \int \frac{1}{N} \sum_{n=1}^N q_{\phi}(z|x_n) \ln q_{\phi}(z|x_n) dz$$

$$= \int q_{\phi}(z) \ln P_{\lambda}(z) dz - \int \sum_{n=1}^N \frac{1}{N} q_{\phi}(z|x_n) \ln q_{\phi}(z|x_n) dz$$

$$= -CE[q_{\phi}(z) \parallel P_{\lambda}(z)] + H[q_{\phi}(z|x)]$$

$$P_{\theta}(x'|z') \sim N(\mu, I) = \frac{1}{I\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x'-\mu}{I}\right)^2} \quad (۱۳)$$

$$\Rightarrow E_{z' \sim q_{\theta}(z|x')} [\lg P_{\theta}(x'|z')]$$

$$= E_{z' \sim q_{\theta}(z|x')} \left[\lg \frac{1}{I\sqrt{2\pi}} - \frac{1}{2} \left(\frac{x'-\mu}{I} \right)^2 \right]$$

$$= E_{z' \sim q_{\theta}(z|x')} \left[-\frac{1}{2} \lg 2\pi - \underbrace{\frac{1}{2} \left(\frac{x'-\mu}{I} \right)^2}_{MSE} \right]$$

$$P_{\theta}(x'|z') \sim \text{bernouly}(a_i) = \prod_{i=1}^d a_i^{x_i} (1-a_i)^{1-x_i} \quad (۱۴)$$

$$\Rightarrow E_{z' \sim q_{\theta}(z|x')} [\lg P_{\theta}(x'|z')]$$

$$= E_{z' \sim q_{\theta}(z|x')} \left[\underbrace{\sum_{i=1}^d (x_i \lg a_i + (1-x_i) \lg (1-a_i))}_{BCE} \right]$$

ج) از آنجایی که BCE برای حل مسائل دسته‌بندی کاربرد دارد، اگر بخواهیم در اینجا background و اجزاء را جدا کنیم یک دسته‌بندی داریم و می‌توان از BCE استفاده کرد. اما اگر هدف این باشد که تصاویر ورودی را بازسازی کنیم بهتر است از MSE استفاده کرد، در این صورت در طول زمان فاصله تصاویر بازسازی شده از تصویر اصلی کمتر می‌شود. همچنین از آنجایی که تابع MSE متقارن و مثبت است می‌توانیم از آن برای آموزش شبکه‌های تولیدی و ورودی را جدا کنیم BCE متقارن نیست و مثبت می‌پذیرد و متقارن نیست، بسیار مناسبی برای نزدیک کردن دو تصویر تولیدی و ورودی نیست. همچنین در مسائل unsupervised بدون label نداریم می‌توان از BCE استفاده کرد.

بازن $\theta \neq 0$ داریم: (۲) (۴)

$$D_{KL}(P||Q) = \sum_{\substack{x=0 \\ y \sim U(0,1)}} P(x,y) \lg \frac{P(x,y)}{Q(x,y)}$$

استعلام x, y →

$$= \sum_{\substack{x=0 \\ y \sim U(0,1)}} P(x=0) P(y) \lg \frac{P(x=0) P(y)}{Q(x=0) Q(y)} = \sum_{\substack{x=0 \\ y \sim U(0,1)}} 1 \cdot \lg \frac{1}{1} = +\infty$$

$$D_{KL}(Q||P) = \sum_{\substack{x=0 \\ y \sim U(0,1)}} Q(x,y) \lg \frac{Q(x,y)}{P(x,y)}$$

استعلام x, y →

$$= \sum_{\substack{x=0 \\ y \sim U(0,1)}} Q(x=0) Q(y) \lg \frac{Q(x=0) Q(y)}{P(x=0) P(y)} = \sum_{\substack{x=0 \\ y \sim U(0,1)}} 1 \cdot \lg \frac{1}{1} = +\infty$$

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}\left(P || \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q || \frac{P+Q}{2}\right)$$

$$= \frac{1}{2} \left(\sum_{\substack{x=0 \\ y \sim U(0,1)}} 1 \cdot \lg \frac{1}{1/2} + \sum_{\substack{x=0 \\ y \sim U(0,1)}} 1 \cdot \lg \frac{1}{1/2} \right) = \lg 2 \quad W(P, Q) = 101$$

اما اگر $\theta = 0$ باشد، این دو توزیع کاملاً منطبق بر هم هستند بنابراین داریم:

$$D_{KL}(P||Q) = 0 \quad D_{KL}(Q||P) = 0 \quad D_{JS}(P, Q) = 0 \quad W(P, Q) = 0$$

(ب) اگر بخواهیم JSD دو توزیع مستقل نابینا باشد، باید یک توزیع در یک نقطه احتمال یک

و دیگری نقاط همزمان داشته باشد. در این صورت توزیع دیگری می‌تواند هم‌بسته باشد. مثلاً

می‌توان P را یک توزیع به شکل تابع دلتا باشد

(ج) خاصیتی که JSD به ازای $\theta = 0$ و از هر دو به ازای $\theta \neq 0$ برای ثابت $\lg 2$ می‌باشد.

در واقع JSD دارای یک ویژگی ناگهانی در $\theta = 0$ است بنابراین در این نقطه مستقل نابینا است.

این مشکل در معیار فاصله W وجود ندارد و این معیار یک مقدار پیوسته به ازای θ های مختلف دارد.

بنابراین این معیار برای آموزش GAN مناسب‌تر است.

(۵) (۲) فردی Discriminator در GAN یک ادیان است که نشان دهنده میزان واقعی بودن نمونه دردی را نشان می دهد. همچنین با استفاده از binary cross entropy آموزش می بیند. این درحالی است که واحد critic در WGAN به گونه ای آموزش می بیند که فاصله Wasserstein بین توزیع های واقعی و غیر واقعی را تخمین بزند. این واحد با استفاده از تابع هزینه ای آموزش می بیند که فردی critic را بجای داده های واقعی مقداری بالا و بجای داده های غیر واقعی مقداری کم ارائه دهد.

(ب) همانند گفته شده می توانیم که فردی critic و فلاف فردی discriminator در یک بازه محدود قرار بگیرد. فردی این واحد باید فاصله Wasserstein بین دو توزیع واقعی و غیر واقعی را تخمین بزند. تابع فعال سازی sigmoid فردی را بین بازه (۰، ۱) نگاشت می کند، بنابراین مناسب است. از طرفی به علت داشتن مشتق صفر در دردی های صلی و صلی یا صلی، باعث ایجاد gradient vanishing می شود که باعث مشکل کردن فرآیند آموزش می شود.

(ج) شبکه critic باید فاصله Wasserstein بین توزیع نمونه های واقعی و غیر واقعی را تخمین بزند. آموزش دقیق و مناسب این شبکه از آموزش شبکه مولد مشکل دار است، همچنین اگر این شبکه بتواند کار خود را به خوبی انجام دهد، می تواند باعث شود که شبکه مولد نیز نمونه های بهتری تولید کند. از محدودیت هایی که برای آموزش وزن های critic داریم این است که این وزن ها باید از طریق Wasserstein loss و روش gradient penalty آپدیت شوند. علاوه بر این وزن ها باید از طریق تکنیک weight clipping آپدیت شوند. در این تکنیک وزن ها در یک بازه کوچک مقید می شوند. این کار تضمین می کند که فردی critic به ازای تغییرات کوچک در دردی، تغییر زیادی نداشته باشد.