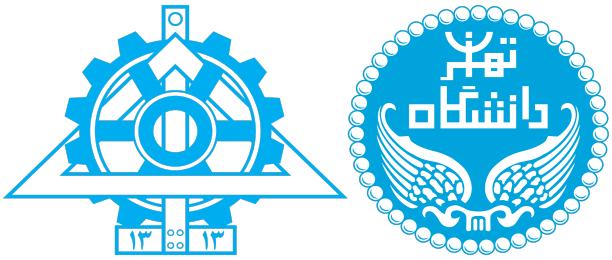


دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

تمرین شماره: ۱

مدل‌های مولد عمیق

نام و نام‌خانوادگی: پریسا محمدی

شماره دانشجویی: ۸۱۰۱۰۱۵۰۹

نیم‌سال اول

سال تحصیلی ۴۰۵-۴۰۴

فهرست مطالب

فهرست تصاویر

بخش اول

زیربخش اول : رسم شبکه بیزی

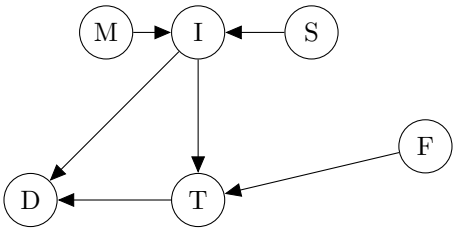
توضیح

برای رسم این گراف، باید روابط علت و معلولی که در متن توضیح داده شده را پیدا کنیم. ما به دنبال این هستیم که کدام متغیر والد متغیر دیگری است

- گره‌های ریشه: متغیرهای M (قدرت سیستم ایمنی)، S (فصل) و F (توانایی مالی) ریشه‌ها هستند. در متن، عاملی بر روی آن‌ها تأثیرگذار نیست، بلکه آن‌ها بر سایر متغیرها تأثیر می‌گذارند.
- گره I: شدت بیماری هم تحت تأثیر قدرت سیستم ایمنی (M) است و هم فصل (S). پس M و S والدین I هستند.
- گره T: انتخاب دارو به شدت بیماری (I) بستگی دارد. همچنین توانایی مالی (F) فرد نیز در انتخاب دارو مؤثر است. پس I و F والدین T هستند.
- گره D: احتمال مرگ، به شدت بیماری (I) و نوع داروی مصرفی (T) بستگی دارد.

رسم گراف

بر اساس تحلیل بالا، گراف شبکه بیزی به صورت زیر رسم می‌شود.



شکل ۱: ساختار شبکه بیزی بر اساس توضیحات

زیربخش دوم : توزیع احتمالأل مشترک

برای نوشتن توزیع احتمال مشترک بر اساس شبکه بیزی، از قاعده زنجیره‌ای استفاده می‌کنیم. این قانون می‌گوید که توزیع توأم، برابر است با حاصل‌ضرب احتمال هر متغیر به شرط والدینش.

والدین هر گره در گراف به این صورت هستند:

- والدین M: ندارد (ریشه)
- والدین S: ندارد (ریشه)
- والدین F: ندارد (ریشه)
- والدین I: M و S
- والدین T: I و F
- والدین D: I و T

بنابراین، توزیع احتمال مشترک به صورت زیر نوشته می‌شود:

$$P(M, S, F, I, T, D) = P(M) \cdot P(S) \cdot P(F) \cdot P(I \mid M, S) \cdot P(T \mid I, F) \cdot P(D \mid I, T)$$

زیربخش سوم (۱۰ نمره): بررسی درستی عبارت‌ها

در این بخش، با استفاده از مفهوم d-separation در گراف، درستی یا نادرستی هر عبارت را بررسی می‌کنیم:

۱. $F \perp D$: نادرست.

دلیل: یک مسیر به صورت $F \to T \to D$ وجود دارد. چون گره میانی T مشاهده نشده است، این مسیر باز است و اطلاعات می‌تواند منتقل شود. پس این دو متغیر مستقل نیستند.

۲. $S \perp D \mid I$: درست.

دلیل: ما باید تمام مسیرهای بین S و D را با فرض مشاهده شدن I بررسی کنیم. مسیر اول $S \to I \to D$ است. چون گره میانی I مشاهده شده، این مسیر زنجیره‌ای مسدود است. مسیر دوم $S \to I \to T \to D$ است. چون گره I مشاهده شده، این مسیر هم مسدود است. چون تمام مسیرها مسدود می‌شوند، این دو متغیر به شرط I مستقل هستند.

۳. $M \perp F$: درست.

دلیل: تنها مسیر بین M و F به صورت $M \to I \to T \leftarrow F$ است. این مسیر یک ساختار برخوردکننده در گره T دارد. تا زمانی که T یا فرزندانش مشاهده نشده باشند، این مسیر مسدود است. چون در شرط چیزی مشاهده نشده، مسیر مسدود است و دو متغیر مستقل هستند.

۴. $M \perp F \mid T$: نادرست.

دلیل: این برعکس حالت قبل است. مسیر $M \to I \to T \leftarrow F$ را داریم. در اینجا، ما خود گره برخوردکننده (T) را مشاهده کرده‌ایم. مشاهده گره برخوردکننده، مسیر را باز می‌کند. چون مسیر باز است، این دو متغیر مستقل نیستند.

۵. $M \perp T \mid \{D, I\}$: نادرست.

دلیل: ما مسیرهای بین M و T را با فرض مشاهده شدن D و I بررسی می‌کنیم. مسیر اول $M \to I \to T$ است. چون گره میانی I مشاهده شده، این مسیر مسدود است. اما مسیر دوم $M \to I \to D \leftarrow T$ است. این مسیر یک ساختار برخوردکننده در گره D دارد. در شرط، گره D (که همان گره برخوردکننده است) مشاهده شده است. مشاهده گره برخوردکننده این مسیر را باز می‌کند. چون حداقل یک مسیر باز وجود دارد، این دو متغیر مستقل نیستند.

بخش دوم

زیربخش اول : توزیع احتمال مشترک

برای نوشتن توزیع احتمال مشترک، از همان قاعده زنجیره‌ای شبکه‌های بیزی استفاده می‌کنیم، یعنی حاصل‌ضرب احتمال هر گره به شرط والدینش.

ابتدا والدین هر گره را از روی گراف مشخص می‌کنیم:

- والدین C: ندارد
- والدین O: ندارد
- والدین A: ندارد
- والدین S: O
- والدین T: O، A
- والدین B: S
- والدین M: A، B، T

با ضرب کردن همه‌ی این احتمالات شرطی، توزیع احتمال توأم به صورت زیر به دست می‌آید:

$$P(C, O, A, S, T, B, M) = P(C)P(O)P(A)P(S \mid O)P(T \mid O, A)P(B \mid S)P(M \mid A, B, T)$$

زیربخش دوم: Markov Blanket متغیر T

مارکوف بلنکت یک گره، مجموعه‌ای از گره‌هایی است که آن گره را از بقیه شبکه مستقل (d-separate) می‌کند.

برای پیدا کردن مارکوف بلنکت گره T، باید سه گروه از گره‌ها را پیدا کنیم:

۱. والدین گره T: { O، A }
۲. فرزندان گره T: { M }
۳. سایر والدین فرزندان گره T (که به آن‌ها «همسر» هم گفته می‌شود): گره M (فرزند T)، والدین دیگری به جز T هم دارد که عبارتند از { A، B }.

با اجتماع گرفتن از همه‌ی این گره‌ها (بدون تکرار)، مارکوف بلنکت گره T به دست می‌آید:

$$\mathbf{MB}(T) = \{O, A, M, B\}$$

زیربخش سوم : Perfect I-Map

خیر، این گراف مارکوف یک Perfect I-Map برای گراف بیزی مربوطه نیست.

چون یک Perfect I-Map باید دقیقاً همان مجموعه‌ی استقلال‌های آماری را نشان دهد که گراف بیزی اصلی نشان می‌دهد.

در اینجا یک تناقض وجود دارد:

- در گراف بیزی (بخش قبل): گره‌های O و A از هم مستقل بودند ($O \perp A$). دلیلش این بود که هر دو ریشه بودند و تنها مسیر بین آن‌ها ($O \rightarrow T \leftarrow A$) توسط یک برخوردکننده در T مسدود شده بود.
- در گراف مارکوف (این بخش): گره‌های O و A مستقیماً با یک یال به هم وصل هستند..

نتیجه: گراف بیزی می‌گوید O و A مستقل هستند، اما گراف مارکوف می‌گوید به هم وابسته‌اند. چون این دو گراف در مورد حداقل یک استقلال با هم اختلاف نظر دارند، این گراف مارکوف یک Perfect I-Map نیست.

زیربخش چهارم : گراف Chordal

خیر، این گراف Chordal (وتری) نیست.

چون:

- تعریف: یک گراف Chordal گرافی است که در آن هر «دور» با طول ۴ یا بیشتر، حداقل یک «وتر» (chord) داشته باشد. (وتر، یالی است که بین دو گره غیرمجاور در آن دور قرار دارد).
- مثال نقض: در این گراف، یک دور به طول ۴ به صورت $O - S - B - T - O$ وجود دارد.
- گره‌های غیرمجاور در این دور، جفت‌های (O، B) و (S، T) هستند.
- همانطور که در گراف می‌بینید، هیچ یالی بین O و B و همچنین هیچ یالی بین S و T وجود ندارد.

نتیجه: از آنجایی که ما یک دور به طول ۴ پیدا کردیم که هیچ وتری ندارد، این گراف Chordal نیست.

زیربخش پنجم : توزیع احتمال مشترک بر اساس maximal cliques

بر اساس قضیه Hammersley-Clifford، توزیع احتمال مشترک در یک گراف مارکوف، متناسب است با حاصل‌ضرب توابع پتانسیل (ϕ) روی تمام کلیک‌های ماکسیمال گراف.

$$P(X) = \frac{1}{Z} \prod_{k} \phi_k(C_k)$$

که در آن Z ثابت نرمال‌سازی است.

ابتدا کلیک‌های ماکسیمال گراف مارکوف (از زیربخش سوم) را پیدا می‌کنیم:

- گره تنها C: { C }
- مثلث‌های موجود در گراف: { O، A، T } و { A، M، T } و { B، M، T }
- یال‌هایی که بخشی از مثلث نیستند اما نمی‌توان گرهی به آن‌ها اضافه کرد: { O، S } و { S، B }

بنابراین، توزیع احتمال مشترک به صورت زیر نوشته می‌شود:

$$P(A,\ldots,T) = \frac{1}{Z}\cdot\phi_1(C)\cdot\phi_2(O,A,T)\cdot\phi_3(A,M,T)\cdot\phi_4(B,M,T)\cdot\phi_5(O,S)\cdot\phi_6(S,B)$$

زیربخش ششم : درستی دو عبارت

می‌خواهیم نشان دهیم که چرا حذف متغیر تنها C در دو حالت متفاوت است.

در حالت گراف بیزی

در گراف بیزی، چون C یک گره تنها (مستقل از بقیه) است، توزیع مشترک به سادگی در آن ضرب می‌شود:

$$P(\text{all}) = P(C)\cdot P(\text{rest\_of\_variables})$$

برای حذف C (به دست آوردن توزیع حاشیه‌ای)، از $\mathrm{P(all)}$ روی C انتگرال می‌گیریم:

$$P(\text{rest}) = \int P(C)\cdot P(\text{rest\_of\_variables})dC$$

چون $P(\text{rest\_of\_variables})$ به C وابسته نیست، از انتگرال خارج می‌شود:

$$P(\text{rest}) = P(\text{rest\_of\_variables})\cdot\int P(C)dC$$

در یک شبکه بیزی، $P(C)$ یک توزیع احتمال است و بنا به تعریف، انتگرال (یا جمع) روی کل دامنه آن باید برابر ۱ باشد:

$$\int P(C)dC = 1$$

بنابراین، فاکتور $P(C)$ به سادگی حذف می‌شود.

در حالت گراف مارکوف

در گراف مارکوف، توزیع مشترک بر اساس پتانسیل‌ها نوشته می‌شود:

$$P(\text{all}) = \frac{1}{Z}\cdot\phi(C)\cdot\phi(\text{rest\_of\_factors})$$

برای حذف C، انتگرال می‌گیریم:

$$P(\text{rest}) = \int \frac{1}{Z} \cdot \phi(C) \cdot \phi(\text{rest\_of\_factors}) dC$$

$$P(\text{rest}) = \frac{1}{Z} \cdot \phi(\text{rest\_of\_factors}) \cdot \int \phi(C) dC$$

در اینجا، $\phi(C)$ یک تابع پتانسیل است، نه لزوماً یک توزیع احتمال. هیچ تضمینی وجود ندارد که $\int \phi(C)dC = 1$ باشد. اما، سوال به ما یک شرط داده است: «...در صورتی که $\int_{-\infty}^{+\infty} \phi(C)dC = 1$». اگر این شرط برقرار باشد، آنگاه می‌توانیم مقدار انتگرال را با ۱ جایگزین کنیم:

$$P(\text{rest}) = \frac{1}{Z} \cdot \phi(\text{rest\_of\_factors}) \cdot (1)$$

بنابراین، تنها تحت آن شرط خاص، می‌توان فاکتور $\phi(C)$ را حذف کرد.

بخش سوم

زیربخش اول : توزیع احتمال مشترک بر اساس maximal cliques

برای نوشتن توزیع احتمال مشترک در یک گراف مارکوف، از قضیه Hammersley-Clifford استفاده می‌کنیم که می‌گوید توزیع، متناسب با حاصل‌ضرب توابع پتانسیل (ϕ) روی «کلیک‌های ماکسیمال» است.

ابتدا کلیک‌های ماکسیمال (گروه‌هایی که همه‌ی اعضایشان به هم وصل هستند و نمی‌توان گرهی به آن‌ها افزود) را پیدا می‌کنیم:

- { A، B، C } (مثلث)
- { B، D، E } (مثلث)
- { C، D، F } (مثلث)
- { E، G } (یال)

حالا توزیع مشترک را بر اساس این کلیک‌ها می‌نویسیم:

$$P(A,B,C,D,E,F,G) = \frac{1}{Z} \cdot \phi_1(A,B,C) \cdot \phi_2(B,D,E) \cdot \phi_3(C,D,F) \cdot \phi_4(E,G)$$

که در آن Z ثابت نرمال‌سازی است.

زیربخش دوم : درستی یا نادرستی موارد

در گراف مارکوف، برای بررسی استقلال $X \perp Y \mid Z$ ، باید ببینیم آیا مجموعه‌ی Z تمام مسیرها را بین X و Y مسدود می‌کند یا خیر.

۱. $G \perp A$: نادرست.

دلیل: برای استقلال کامل (بدون شرط)، نباید هیچ مسیری بین دو گره وجود داشته باشد. اما مسیرهای باز زیادی مانند $G - E - B - A$ وجود دارد. چون مسیر باز هست، دو گره مستقل نیستند.

۲. $F \perp A \mid \{D, C\}$: درست.

دلیل: باید بررسی کنیم که آیا مجموعه‌ی { C، D } تمام مسیرهای بین F و A را مسدود می‌کند. مسیر $F - C - A$ توسط گره C (که در شرط آمده) مسدود می‌شود. مسیر $F - D - B - A$ توسط گره D (که در شرط آمده) مسدود می‌شود. تمام مسیرهای دیگر نیز (مانند $F - D - E - B - A$) از حداقل یکی از این دو گره عبور می‌کنند. چون همه مسیرها مسدود هستند، استقلال برقرار است.

۳. $G \perp C \mid E$: درست.

دلیل: باید بررسی کنیم که آیا { E } مسیرهای G به C را مسدود می‌کند. گره G فقط به گره E متصل است. بنابراین، هر مسیری از G به C (مانند $G - E - B - C$ یا $G - E - D - C$) باید از E عبور کند. چون E در مجموعه‌ی شرطی است (مشاهده شده)، تمام مسیرهای ممکن از G به C در همان گام اول مسدود می‌شوند. پس استقلال برقرار است.

۴. $p(A \mid B, C) = p(A \mid B, C, E)$: درست.

دلیل: این عبارت، شکل دیگری از بیانِ استقلال $A \perp E \mid \{B, C\}$ است. یعنی می‌پرسد، آیا A و E به شرط B و C ، مستقل هستند؟ باید بررسی کنیم که آیا { B، C } تمام مسیرهای A به E را مسدود می‌کند. مسیر $A - B - E$ توسط گره B مسدود می‌شود. مسیر $A - C - D - E$ توسط گره C مسدود می‌شود. مسیر $A - B - D - E$ نیز توسط گره B مسدود می‌شود. چون تمام مسیرها بین A و E توسط مجموعه‌ی { B، C } مسدود می‌شوند، این دو به شرط هم مستقل هستند و عبارت داده شده درست است.

زیربخش سوم : تغییر تابع پتانسیل

اگر تابع پتانسیل $\phi(E,G)$ را ۵ برابر کنیم، توزیع احتمال نهایی هیچ تغییری نمی‌کند.

دلیل این موضوع، ثابت نرمال‌سازی (Z) است.

- توزیع احتمال اولیه: $P_{\text{old}} = \frac{1}{Z_{\text{old}}} \cdot \phi_{\text{rest}} \cdot \phi(E,G)$
- توزیع جدید (قبل از نرمال‌سازی): $\tilde{P}_{\text{new}} = \phi_{\text{rest}} \cdot (5 \cdot \phi(E,G))$
- ثابت نرمال‌سازی جدید: $Z_{\text{new}} = \sum_X \tilde{P}_{\text{new}} = \sum_X [5 \cdot \phi_{\text{rest}} \cdot \phi(E,G)]$

$$Z_{\text{new}} = 5 \cdot [\sum_X \phi_{\text{rest}} \cdot \phi(E,G)] = 5 \cdot Z_{\text{old}}$$

- توزیع احتمال نهایی: $P_{\text{new}} = \frac{\tilde{P}_{\text{new}}}{Z_{\text{new}}} = \frac{5 \cdot (\phi_{\text{rest}} \cdot \phi(E,G))}{5 \cdot Z_{\text{old}}}$

همانطور که مشخص است، عدد ۵ از صورت و مخرج ساده می‌شود و توزیع نهایی برابر با توزیع اولیه خواهد بود:

$$P_{\text{new}} = \frac{\phi_{\text{rest}} \cdot \phi(E,G)}{Z_{\text{old}}} = P_{\text{old}}$$

بخش چهارم

پیدا کردن پارامتر بهینه θ

هدف ما در (Inference Variational)، به حداقل رساندن واگرایی KL بین توزیع تقریبی $q(z)$ و توزیع پسین $p(z \mid x)$ است. این کار معادلِ ماکسیمم کردن معیار ELBO است.

گام ۱: نوشتن فرمول ELBO

فرمول کلی ELBO به صورت زیر تعریف می‌شود:

$$\mathcal{L}(\theta) = E_q[\log p(x, z)] - E_q[\log q(z)]$$

ما باید هر کدام از این دو عبارت امید ریاضی را محاسبه کنیم.

گام ۲: محاسبه بخش اول $E_q[\log p(x, z)]$

ابتدا $\log p(x, z)$ را با استفاده از $p(x, z) = p(x \mid z)p(z)$ محاسبه می‌کنیم.

$$\log p(z) = \log(e^{-z}) = -z$$

$$\log p(x \mid z) = \log(ze^{-zx}) = \log(z) + \log(e^{-zx}) = \log(z) - zx$$

با جمع کردن این دو، به $\log p(x,z)$ می‌رسیم:

$$\log p(x,z) = \log(z) - zx - z = \log(z) - z(x+1)$$

حالا از این عبارت نسبت به توزیع q امید ریاضی می‌گیریم:

$$E_q[\log p(x,z)] = E_q[\log(z) - z(x+1)]$$

$$E_q[\log p(x,z)] = E_q[\log(z)] - E_q[z(x+1)]$$

چون x ثابت است، $(x+1)$ از امید ریاضی بیرون می‌آید:

$$E_q[\log p(x,z)] = E_q[\log(z)] - (x+1)E_q[z]$$

در نهایت، مقدار داده شده $E_q[z] = \frac{2}{\theta}$ را جایگذاری می‌کنیم:

$$E_q[\log p(x,z)] = E_q[\log(z)] - (x+1)\frac{2}{\theta}$$

گام ۳: محاسبه بخش دوم $E_q[\log q(z)]$

این بخش، منفیِ آنتروپی توزیع q است. ابتدا $\log q(z)$ را محاسبه می‌کنیم:

$$\log q(z) = \log(\theta^2 z e^{-\theta z}) = \log(\theta^2) + \log(z) + \log(e^{-\theta z})$$

$$\log q(z) = 2\log(\theta) + \log(z) - \theta z$$

حالا از این عبارت نسبت به q امید ریاضی می‌گیریم:

$$E_q[\log q(z)] = E_q[2\log(\theta) + \log(z) - \theta z]$$

$$E_q[\log q(z)] = 2\log(\theta) + E_q[\log(z)] - \theta E_q[z]$$

دوباره $E_q[z] = \frac{2}{\theta}$ را جایگذاری می‌کنیم:

$$E_q[\log q(z)] = 2\log(\theta) + E_q[\log(z)] - \theta\left(\frac{2}{\theta}\right)$$

$$E_q[\log q(z)] = 2\log(\theta) + E_q[\log(z)] - 2$$

گام ۴: تشکیل ELBO نهایی و ساده‌سازی

حالا دو بخش محاسبه شده را در فرمول اصلی ELBO قرار می‌دهیم:

$$\mathcal{L}(\theta) = E_q[\log p(x,z)] - E_q[\log q(z)]$$

$$\mathcal{L}(\theta) = \left(E_q[\log(z)] - \frac{2(x+1)}{\theta}\right) - (2\log(\theta) + E_q[\log(z)] - 2)$$

با باز کردن پرانتز، عبارت $E_q[\log(z)]$ (که محاسبه‌ی سختی داشت) از طرفین حذف می‌شود:

$$\mathcal{L}(\theta) = E_q[\log(z)] - \frac{2(x+1)}{\theta} - 2\log(\theta) - E_q[\log(z)] + 2$$

$$\mathcal{L}(\theta) = -\frac{2(x+1)}{\theta} - 2\log(\theta) + 2$$

گام ۵: ماکسیمم کردن ELBO و پیدا کردن θ

برای پیدا کردن θ بهینه، از $\mathcal{L}(\theta)$ نسبت به θ مشتق می‌گیریم و برابر صفر قرار می‌دهیم.

$$\frac{d\mathcal{L}}{d\theta} = \frac{d}{d\theta}\left(-2(x+1)\theta^{-1} - 2\log(\theta) + 2\right)$$

$$\frac{d\mathcal{L}}{d\theta} = (-2(x+1)) \cdot (-1 \cdot \theta^{-2}) - 2 \cdot \left(\frac{1}{\theta}\right) + 0$$

$$\frac{d\mathcal{L}}{d\theta} = \frac{2(x+1)}{\theta^2} - \frac{2}{\theta}$$

حالا مشتق را برابر صفر می‌گذاریم:

$$\frac{2(x+1)}{\theta^2} - \frac{2}{\theta} = 0$$

$$\frac{2(x+1)}{\theta^2} = \frac{2}{\theta}$$

با ساده کردن 2 از طرفین و $1/\theta$ از مخرج‌ها، به این می‌رسیم:

$$\frac{x+1}{\theta} = 1$$

$$\theta = x + 1$$

پاسخ نهایی

مقدار بهینه پارامتر θ برابر با $x + 1$ است.

بخش پنجم: پیاده‌سازی VAE

زیر بخش اول : تابع هزینه VAE

این سوال دو بخش دارد: ۱) چرا مستقیماً $p(x)$ را بهینه نمی‌کنیم؟ ۲) این عبارت (ELBO) چطور مدل را بهینه می‌کند؟

۱. چرا مستقیماً $p(x)$ را بهینه نمی‌کنیم؟

پاسخ کوتاه: چون محاسبه‌ی مستقیم $p(x)$ **«محاسبه‌ناپذیر» (intractable)** است.

توضیح: برای به دست آوردن $p(x)$، ما باید متغیر پنهان z را به حاشیه برانیم. یعنی باید این انتگرال را حساب کنیم: $p(x) = \int p(x|z)p(z)dz$. در یک مدل پیچیده مثل VAE، فضای پنهان z ابعاد زیادی دارد. محاسبه‌ی این انتگرال روی تمام مقادیر ممکن z (که یک فضای پیوسته و بسیار بزرگ است) در عمل غیرممکن است. ما نمی‌توانیم یک جواب سرراست برای این انتگرال پیدا کنیم، برای همین می‌گوییم «محاسبه‌ناپذیر» است.

۲. این عبارت (ELBO) چگونه مدل را بهینه می‌کند؟

چون نمی‌توانیم $\log p(x)$ را مستقیماً ماکسیمم کنیم، به سراغ یک حد پایین برای آن می‌رویم و سعی می‌کنیم آن حد پایین را ماکسیمم کنیم. عبارت (۱) همان **ELBO** است.

ریاضیات نشان می‌دهد که $\log p(x) \geq \text{ELBO}$ است. به جای ماکسیمم کردن $\log p(x)$ (که سخت بود)، ما ELBO را ماکسیمم می‌کنیم. با بالا بردن این حد پایین، به طور غیرمستقیم خود $\log p(x)$ را هم تا جای ممکن بالا می‌بریم.

این عبارت از دو بخش تشکیل شده که یک توازن (Trade-off) ایجاد می‌کنند:

۱. **بخش بازسازی (Reconstruction):** $E_{q(z|x)}[\log p(x|z)]$

این بخش همان «خطای بازسازی» است. $q(z|x)$ همان Encoder و $p(x|z)$ همان Decoder است. $\log p(x|z)$ میزان موفقیت Decoder در بازسازی x را اندازه‌گیری می‌کند.

هدف: ماکسیمم کردن این عبارت، که معادل **کمینه کردن** تفاوت بین تصویر ورودی و تصویر بازسازی‌شده است.

۲. **بخش تنظیم‌کننده (Regularization):** $-D_{KL}(q(z|x)||p(z))$

این بخش یک جریمه (Penalty) است. $p(z)$ توزیع «پیشین» (Prior) ما برای z است (معمولاً یک گوسی استاندارد $N(0, I)$). $q(z|x)$ توزیع خروجی Encoder است. KL فاصله بین این دو توزیع را می‌سنجد.

هدف: ما می‌خواهیم D_{KL} را **کمینه** کنیم (چون علامت منفی پشت آن است، در کل عبارت ماکسیمم می‌شود). این کار Encoder را مجبور می‌کند که فضای پنهان را ساده و منظم (شبیه به $N(0, I)$) نگه دارد.

خلاصه: VAE با ماکسیمم کردن ELBO یاد می‌گیرد که همزمان هم خوب بازسازی کند و هم فضای پنهان را منظم نگه دارد.

زیر بخش دوم : توضیح داده dsprites

مجموعه داده dsprites یک مجموعه داده معروف و مصنوعی (رندر شده) است که به طور خاص برای تحقیق در مورد بازنمایی‌های گسسته در مدل‌های مولد ساخته شده است.

- **مشخصات کلی داده:**
  - **محتوا:** تصاویر سیاه و سفید (باینری) از ۳ شکل ساده: **مربع، بیضی (دایره)، و قلب**.
  - **ابعاد تصویر:** هر تصویر 64x64 پیکسل است.
  - **تعداد کل تصاویر:** 737,280. این عدد از ترکیب تمام حالات ممکن عوامل نهفته به دست آمده است (۳ شکل × ۶ مقیاس × ۴۰ چرخش × ۳۲ موقعیت افقی × ۳۲ موقعیت عمودی).

- **حجم داده:** هر تصویر 64x64 است. در فایل npz.، این تصاویر به صورت باینری (یا uint8) ذخیره شده‌اند که هر تصویر 4KB (64*64*1 بایت) و کل داده (بدون فشرده‌سازی) حدود 3GB حجم دارد.

• **عوامل نهفته :** هر تصویر دقیقاً بر اساس چند عامل نهفته مشخص و مستقل از هم ساخته شده است:

- شکل (۳ نوع)
- مقیاس (۶ اندازه)
- چرخش (۴۰ زاویه)
- موقعیت X (۳۲ موقعیت)
- موقعیت Y (۳۲ موقعیت)

• **مزایای این داده:**

- **وجود عوامل واقعی :(Ground-Truth)** بزرگترین مزیت این داده این است که ما دقیقاً می‌دانیم هر تصویر با چه پارامترهایی ساخته شده. این به ما اجازه می‌دهد تا به طور کمی ارزیابی کنیم که آیا مدل VAE ما توانسته این عوامل را به درستی در فضای پنهان z کشف و جدا (disentangle) کند یا نه.
- **سادگی:** تصاویر باینری و ساده هستند. این باعث می‌شود مدل به جای یادگیری بافت‌های (texture) پیچیده (مانند تصاویر واقعی)، روی یادگیری ساختار و عوامل هندسی تمرکز کند.
- **استاندارد بودن:** این داده به یک معیار استاندارد و پذیرفته‌شده برای مقایسه مدل‌های جذابی پذیر تبدیل شده است.

• **معایب این داده:**

- **سادگی بیش از حد:** این داده هیچ شباهتی به داده‌های دنیای واقعی (که دارای رنگ، بافت، نورپردازی، سایه و چندین شیء درهم‌تنیده هستند) ندارد.
- **استقلال کامل عوامل:** در این داده، تمام عوامل (مثل اندازه و چرخش) کاملاً از هم مستقل هستند. در دنیای واقعی، عوامل اغلب با هم همبستگی دارند (مثلاً در یک داده خودرو، اندازه خودرو و تعداد درها همبستگی دارند). این، وظیفه جداسازی را برای مدل به طور مصنوعی آسان می‌کند.

زیر بخش سوم : ترفند پارامتردهی مجدد

۱. مشکل :

Encoder در VAE به جای یک بردار z، پارامترهای یک توزیع (مثلاً میانگین μ و واریانس σ^2) را خروجی می‌دهد. سپس، ما باید یک z از این توزیع نمونه‌گیری (Sample) کنیم: $z \sim N(\mu, \sigma^2)$.

مشکل اصلی این است که عملیات نمونه‌گیری یک عملیات **تصادفی (Stochastic)** است و مشتق‌پذیر نیست.

این باعث می‌شود که جریان گرادیان در گره نمونه‌گیری قطع شود و ما نتوانیم با backpropagation پارامترهای Encoder (که μ و σ را ساخته‌اند) را آپدیت کنیم.

۲. راه‌حل: ترفند پارامتردهی مجدد

ایده اصلی این است که ما بخش تصادفی را از بخش پارامتری (که می‌خواهیم از آن مشتق بگیریم) جدا کنیم.

ما فرمول نمونه‌گیری را بازنویسی می‌کنیم:

۱. به جای $z \sim N(\mu, \sigma^2)$

۲. ابتدا یک نویز ساده و استاندارد ϵ (اپسیلون) از $N(0, 1)$ نمونه‌گیری می‌کنیم.

۳. سپس، z را به صورت یک محاسبه‌ی «قطعی» (deterministic) با استفاده از μ، σ، و ϵ می‌سازیم:

$$z = \mu + \sigma \cdot \epsilon$$

حالا ϵ منبع تصادفی بودن ما است (که نیازی به backprop ندارد) و z به صورت قطعی (فقط با یک ضرب و یک جمع ساده) به μ و σ متصل است.

چون عملیات ضرب و جمع کاملاً مشتق‌پذیر هستند، گرادیان می‌تواند به راحتی از Decoder به z و سپس به μ و σ عبور کند و به Encoder برسد. این به مدل اجازه می‌دهد به صورت سرتاسری (end-to-end) آموزش ببیند.

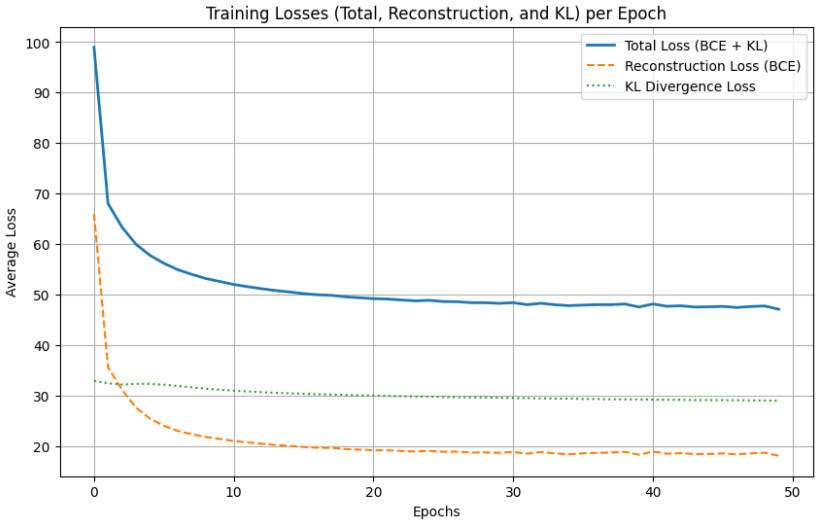
زیر بخش چهارم : بررسی عملکرد مدل

پس از اتمام آموزش مدل با تابع هزینه اصلاح‌شده ،(per-sample) نتایج به دست آمده در نمودار هزینه (شکل ۲) و تصاویر بازسازی شده (شکل ۳) به شرح زیر تحلیل می‌شوند.

۱. تحلیل نمودار هزینه‌ها

نمودار هزینه نشان می‌دهد که آموزش مدل کاملاً موفقیت‌آمیز بوده است:

- **هزینه بازسازی :(BCE)** این هزینه (خط چین نارنجی) در همان اپاک‌های اولیه به شدت کاهش یافته و به مقدار پایداری در حدود ۲۰ رسیده است. این نشان می‌دهد که بخش Decoder مدل به سرعت یاد گرفته است که چگونه تصاویر ورودی را با دقت بالا بازسازی کند.
- **هزینه KL:** هزینه KL (خط نقطه‌چین سبز) در مقداری معنادار (حدود ۳۰) شروع شده و به آرامی کاهش یافته و همگرا شده است. صفر نشدن این هزینه نشان می‌دهد که Encoder ورودی را نادیده نگرفته و فضای پنهان z حاوی اطلاعات مفیدی است.
- **هزینه کل :** این هزینه (خط آبی پررنگ) که مجموع دو هزینه دیگر است، به طور پیوسته کاهش یافته و همگرا شده است که نشان‌دهنده یک فرآیند یادگیری پایدار است.

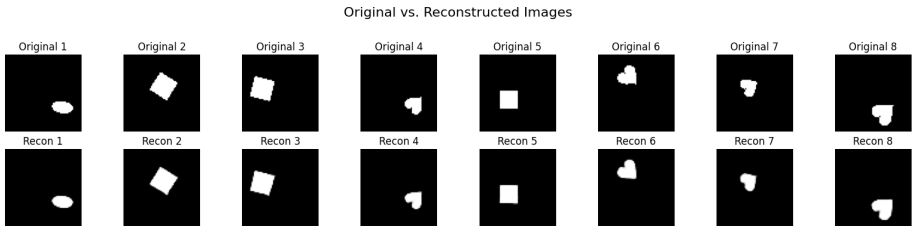


شکل ۲: نمودار هزینه‌های آموزش کل، بازسازی، و KL در طول اپاک‌ها

۲. تحلیل و مقایسه تصاویر

تصاویر بازسازی شده، موفقیت مدل را به وضوح تایید می‌کنند:

- **کیفیت بازسازی:** همانطورکه در شکل ۳ مشاهده می‌شود، ۸ نمونه تصادفی عبور داده شده از مدل، به صورت تقریباً **بی‌نقص و یکسان** بازسازی شده‌اند.
- **یادگیری عوامل:** مدل به وضوح توانسته است تمام عوامل نهفته (factors) دادگان را یاد بگیرد:
  - **شکل:** قلب (نمونه ۴، ۶، ۸)، مربع (نمونه ۲ و ۳) و بیضی (نمونه ۱، ۵، ۷) به درستی تشخیص داده و بازسازی شده‌اند.
  - **موقعیت و چرخش:** مدل به درستی موقعیت و چرخش هر شیء را درک و تکرار کرده است.
- **کیفیت بازسازی :** برخلاف مشکل تاری رایج در بسیاری از VAEها، بازسازی‌های این مدل به دلیل سادگی دادگان، بسیار واضح (sharp) و باکیفیت هستند.



شکل ۳: مقایسه ۸ نمونه تصادفی اصلی (ردیف بالا) و نسخه‌های بازسازی شده (ردیف پایین)

۳. نتیجه‌گیری نهایی

عملکرد مدل با استفاده از تابع هزینه اصلاح‌شده (که مقیاس‌بندی درستی داشت) عالی است. مدل هم توانایی فشرده‌سازی اطلاعات (یادگیری فضای پنهان z) و هم توانایی بازسازی دقیق را به دست آورده و یک توازن سالم بین دو بخش تابع هزینه ELBO برقرار کرده است.

زیر بخش پنجم : توضیح مدل β-VAE

۱. این مدل چه بهبودی دارد؟

مدل β-VAE یک بهبود بسیار مهم نسبت به VAE معمولی دارد: این مدل به ما اجازه می‌دهد که توازن (trade-off) بین کیفیت بازسازی و نظم فضای پنهان را کنترل کنیم.

- مشکل VAE معمولی (که $\beta = 1$ است): در VAE معمولی، مدل به شدت تمایل دارد که «کیفیت بازسازی» را به «نظم فضای پنهان» (بخش KL) ترجیح دهد. این کار اغلب منجر به فروپاشی پسین می‌شود، جایی که D_{KL} به صفر میل می‌کند و z هیچ چیز معناداری یاد نمی‌گیرد.
- راه‌حل β-VAE: با قرار دادن β (بتا) به عنوان یک ضریب برای بخش D_{KL} ما می‌توانیم اهمیت این بخش را تغییر دهیم.
- اگر $\beta > 1$ (مثلاً $\beta = 4$): ما به مدل می‌گوییم که بخش D_{KL} (نظم فضای پنهان)چهار برابر مهم‌تر از بازسازی است. ما مدل را به شدت «جریمه» می‌کنیم اگر $q(z|x)$ (خروجی Encoder) از $p(z)$ (توزیع $N(0,1)$) فاصله بگیرد.
- **نتیجه (بهبود اصلی):** این فشار اضافه، مدل را مجبور می‌کند تا اطلاعات را به **بهینه‌ترین** شکل ممکن در فضای پنهان z فشرده کند. بهینه‌ترین راه برای این کار، این است که هر بُعد z مسئول یک عامل مستقل در داده‌ها (مثلاً z_1 مسئول چرخش، z_2 مسئول اندازه) شود. این همان چیزی است که به آن بازنمایی جداشده می‌گوییم.

۲. آیا رابطه ریاضی آن با نسخه اصلی همخوانی دارد؟

بله، همخوانی دارد. رابطه (۱) (ELBO اصلی) در واقع یک حالت خاص از رابطه (۲) (رابطه β-VAE) است که در آن $\beta = 1$ در نظر گرفته شده است. رابطه (۲) یک فرمول جدید یا متفاوت نیست، بلکه شکل وزن‌دهی شده همان ELBO اصلی است که به ما قابلیت کنترل توازن را می‌دهد.

زیر بخش ششم : آموزش و تحلیل مدل β-VAE

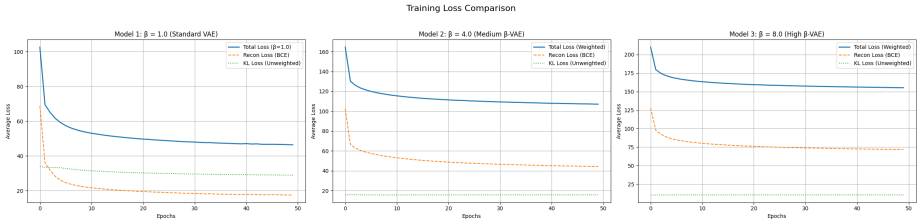
در این بخش، سه مدل با مقادیر مختلف β (۰.۱، ۰.۴ و ۰.۸) آموزش داده شدند. هدف، بررسی تأثیر ضریب β بر توازن بین کیفیت بازسازی و فشار تنظیم‌کنندگی بر روی فضای پنهان است.

۱. تحلیل نمودارهای هزینه

نمودارهای هزینه در شکل ۴، به وضوح این توازن را نشان می‌دهند:

- **مدل $\beta = 1.0$ (VAE استاندارد):** این مدل کمترین هزینه بازسازی (خط نارنجی، حدود ۱۸) را دارد. در مقابل، هزینه KL (خط سبز، حدود ۲۸) آن از همه مدل‌ها بیشتر است. این نشان می‌دهد که مدل، بازسازی دقیق را به نظم فضای پنهان ترجیح داده است و حاضر است با پرداخت هزینه KL بالا، اطلاعات زیادی را در z فشرده کند.

- **مدل** $\beta = 4.0$: با افزایش β به ۰.۴، هزینه بازسازی به طور قابل توجهی افزایش می‌یابد (به حدود ۴۰ می‌رسد). در عوض، مدل مجبور شده است که هزینه KL را کاهش دهد (به حدود ۱۵).
- **مدل** $\beta = 8.0$: با فشار بسیار زیاد $\beta = 8.0$، این توازن به شدت به سمت دیگر متمایل می‌شود. هزینه بازسازی بسیار بالا می‌رود (حدود ۷۵)، که نشان می‌دهد مدل کیفیت بازسازی را فدا کرده است تا بتواند هزینه KL (خط سبز، حدود ۱۸) را تا حد ممکن پایین نگه دارد.

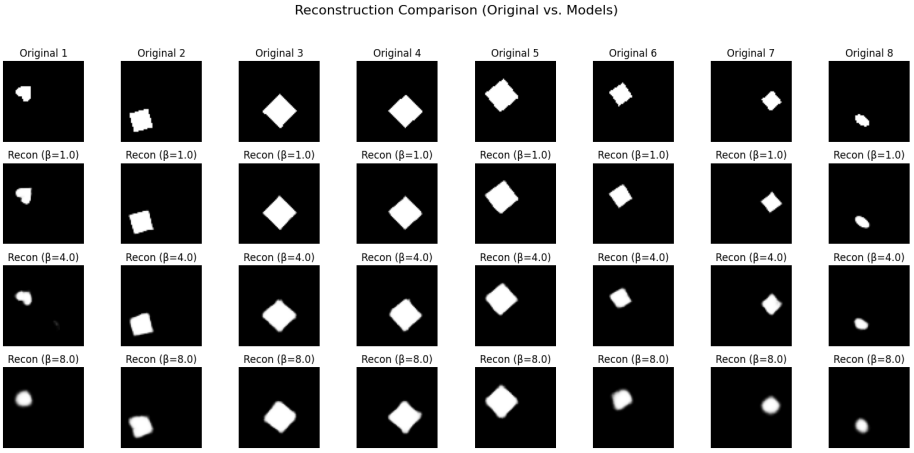


شکل ۴: مقایسه نمودار هزینه‌های آموزش برای مدل‌ها با β مختلف

۲. تحلیل تصاویر بازسازی شده

تصاویر بازسازی شده در شکل ۵، نتایج نمودارهای هزینه را به صورت بصری تأیید می‌کنند:

- **مدل** $\beta = 1.0$: تصاویر بازسازی شده (ردیف دوم) واضح (sharp) و تقریباً یکسان با تصاویر اصلی (ردیف اول) هستند. کیفیت بازسازی عالی است.
- **مدل** $\beta = 4.0$: کیفیت بازسازی کمی افت کرده است. تصاویر (ردیف سوم) کمی تار (blurry) یا لکه (blobbier) شده‌اند. برای مثال، قلب در نمونه ۸، جزئیات دقیق خود را از دست داده است.
- **مدل** $\beta = 8.0$: کیفیت بازسازی به طور واضح قربانی شده است. مدل (ردیف چهارم) دیگر جزئیات را بازسازی نمی‌کند. قلب (نمونه ۶) به یک لکه بیضی‌شکل تبدیل شده است. مربع چرخان (نمونه ۲، ۳، ۴، ۵) به صورت یک مربع صاف و بدون چرخش بازسازی شده است.
- **تحلیل:** این یک یافته بسیار مهم است. مدل تحت فشار $\beta = 8.0$ تشخیص داده که چرخش یک عامل پیچیده است و برای صرفه‌جویی در بودجه KL، تصمیم گرفته است که این اطلاعات را نادیده بگیرد و فقط عامل ساده‌تر (یعنی شکل مربع) را بازسازی کند.



شکل ۵: مقایسه کیفیت بازسازی ۸ نمونه تصادفی در سه مدل مختلف

زیر بخش هفتم (۸ نمره): توضیح و تحلیل معیار MIG

۱. این معیار چیست؟

MIG یک معیار استاندارد برای اندازه‌گیری کمی میزان جداشدگی (Disentanglement) یک مدل است.

- **هدف:** ما می‌خواهیم بدانیم که آیا مدل ما (مثلاً β-VAE) توانسته عوامل مستقل داده (مثل shape، rotation، scale و ...) را در ابعاد مستقل فضای پنهان $(z_1, z_2, ...)$ یاد بگیرد یا نه.
- **ایده اصلی MIG:** یک مدل کاملاً جداشده مدلی است که در آن، یک بُعد پنهان (مثلاً z_3) اطلاعات زیادی در مورد یک عامل واقعی (مثلاً rotation) داشته باشد، و همزمان اطلاعات بسیار کمی در مورد سایر عوامل واقعی (مثل scale یا shape) داشته باشد. MIG دقیقاً همین شکاف (Gap) اطلاعاتی را اندازه‌گیری می‌کند.

۲. چگونه محاسبه می‌شود؟

۱. ما به داده‌های dsprites (که عوامل واقعی k را می‌دانیم) و مدل آموزش‌دیده (که z را به ما می‌دهد) نیاز داریم.

۲. برای **هر بُعد پنهان** z_j (مثلاً z_3) اطلاعات متقابل آن را با تک تک عوامل واقعی k_i (مثل shape، scale، rotation، posX، posY) محاسبه می‌کنیم: $I(z_j; k_{\text{shape}})$، $I(z_j; k_{\text{scale}})$، $I(z_j; k_{\text{rotation}})$، ...

۳. پیدا می‌کنیم که z_j با کدام عامل واقعی **بیشترین** اطلاعات متقابل (MI) را دارد (مثلاً $k_1 = k_{\text{rotation}}$).

۴. پیدا می‌کنیم که با کدام عامل واقعی **دومین** بیشترین اطلاعات متقابل را دارد (مثلاً $k_2 = k_{\text{scale}}$).

۵. شکاف (Gap) را محاسبه می‌کنیم: $I(z_j; k_1) - I(z_j; k_2)$

۶. این شکاف را نرمال‌سازی می‌کنیم (بر آنتروپی $H(k_1)$ تقسیم می‌کنیم).

۳. تحلیل نتایج MIG به دست آمده

در این آزمایش، امتیاز MIG برای هر سه مدل محاسبه شد:

- ($\beta = 1.0$): امتیاز MIG : ۰.۱۶۴۲
- ($\beta = 4.0$): امتیاز MIG : ۰.۱۵۲۹
- ($\beta = 8.0$): امتیاز MIG : ۰.۱۹۸۸

تحلیل امتیازات:

۱. **نتیجه اصلی:** مدل $\beta = 8.0$ به طور قابل توجهی بالاترین امتیاز MIG (۰.۱۹۸۸) را کسب کرده است. این به صورت عددی ثابت می‌کند که فشار (جریمه) سنگین‌تر بر روی KL، مدل را مجبور به یادگیری یک فضای پنهان بسیار ساختاریافته‌تر و جداشده‌تر کرده است.

۲. **تحلیل $\beta = 4.0$:** جالب اینجاست که امتیاز $\beta = 4.0$ (۰.۱۵۲۹) حتی کمی پایین‌تر از $\beta = 1.0$ (۰.۱۶۴۲) است. این نشان می‌دهد که فشار $\beta = 4.0$ هنوز به اندازه کافی قوی نبوده تا مدل را مجبور به جداسازی کند. تفاوت این دو امتیاز ناچیز است و احتمالاً ناشی از نویز تصادفی در آموزش است.

۳. **اهمیت جهش:** چیزی که اهمیت دارد، افت کوچک از ۰.۱ به ۰.۴ نیست، بلکه جهش بزرگ از مقادیر پایین به امتیاز ۰.۱۹۸۸ در $\beta = 8.0$ است. این نشان می‌دهد که یک آستانه بحرانی از فشار β برای دستیابی به جداسازی لازم بوده است.

زیر بخش هشتم: تحلیل PCA

۱. هدف این بخش چیست؟

این بخش یک روش کیفی برای مشاهده ساختار فضای پنهان z است تا ببینیم آیا نتایج بخش هفتم (MIG) را تأیید می‌کند یا نه.

۲. فرآیند کار

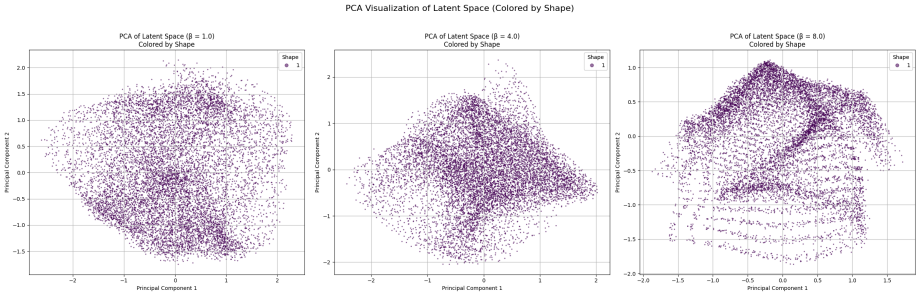
۱. تعداد زیادی نمونه (مثلاً ۵۰۰۰) از دادگان تست را به مدل می‌دهیم و بردارهای پنهان z (که مثلاً ۲۰ بُعدی هستند) را برای همه‌ی آن‌ها استخراج می‌کنیم.

۲. از PCA (تحلیل مؤلفه‌های اصلی) استفاده می‌کنیم تا این فضای ۲۰ بُعدی را به ۲ بُعد (PC1 و PC2) کاهش دهیم.

۳. این ۵۰۰۰ نقطه را در یک نمودار پراکندگی (Scatter Plot) دو بُعدی (PC1 در مقابل PC2) رسم می‌کنیم.

۴. **نکته کلیدی:** ما این نقاط را بر اساس یکی از عوامل واقعی (مثلاً shape یا rotation) رنگ‌آمیزی می‌کنیم.

۳. تحلیل نمودارهای PCA و بررسی نتایج



شکل ۶: مقایسه نمودار PCA فضای پنهان سه مدل (رنگ‌آمیزی بر اساس شکل)

نمودارهای PCA در شکل ۶، نتایج MIG (از بخش هفتم) را به طور کامل تأیید می‌کنند:

- **مدل** $\beta = 1.0$ **و** $\beta = 4.0$: نمودارهای سمت چپ و میانی، هر دو توده‌های درهم و یکنواخت را نشان می‌دهد. رنگ‌ها (که نماینده ۳ شکل مختلف هستند) کاملاً در هم مخلوط شده‌اند. این نشان‌دهنده

یک فضای پنهان درهم‌تنیده است و با امتیازات MIG پایین آن‌ها همخوانی دارد.

- **مدل** $\beta = 8.0$: نمودار سمت راست یک موفقیت کامل را نشان می‌دهد. فضای پنهان دیگر یک توده نیست، بلکه یک شکل ساختاریافته و واضح به خود گرفته است. رنگ‌ها به خوشه‌های کاملاً مجزا تفکیک شده‌اند: هر شاخه (arm) از این شکل، مربوط به یکی از اشکال (قلب، مربع یا بیضی) است. این جداسازی بصری، امتیاز MIG بسیار بالای آن (۰.۱۹۸۸) را تأیید می‌کند.

۴. نتیجه‌گیری نهایی همخوانی

نتایج هر سه بخش (۶، ۷ و ۸) کاملاً با هم همخوانی دارند و توازن β-VAE را به نمایش می‌گذارند:

- $\beta = 1.0$: بهترین بازسازی (BCE پایین، تصاویر واضح)، اما بدترین جداسازی (MIG پایین، PCA درهم‌تنیده).
- $\beta = 8.0$: بدترین بازسازی (BCE بالا، تصاویر تار و با جزئیات اشتباه)، اما بهترین جداسازی (MIG بالا، PCA ساختاریافته).

این آزمایش ثابت می‌کند که با افزایش β (قربانی کردن کیفیت بازسازی)، می‌توانیم مدل را وادار به یادگیری یک فضای پنهان جداشده، معنادار و قابل تفسیر کنیم.

بخش ششم

زیر بخش اول : VQ-VAE

۱. توضیح مدل (گسسته‌سازی)

مدل **VQ-VAE** یک تفاوت کلیدی با VAE معمولی دارد: به جای استفاده از یک «فضای پنهان پیوسته (که با μ و σ تعریف می‌شد)، از یک فضای پنهان گسسته استفاده می‌کند.

این کار به این صورت انجام می‌شود:

۱. **Encoder :** Encoder تصویر ورودی x را می‌گیرد و یک بردار پیوسته (مثلاً z_e) تولید می‌کند.

۲. **Codebook (کتابچه کد):** مدل یک کتابچه کد (e) دارد. این کتابچه شامل K عدد بردار پتانسیل است (مثلاً ۵۱۲ بردار مختلف).

۳. **Quantization (گسسته‌سازی):** مدل، نزدیک‌ترین بردار در کتابچه کد به z_e را پیدا می‌کند.

۴. **خروجی Encoder :** خروجی نهایی، آن بردار نزدیک در کتابچه کد است (مثلاً اندیس شماره ۲۱۰ از بین ۵۱۲ بردار ممکن).

۵. **Decoder:** برای بازسازی، Decoder خودِ بردار پتانسیل» (e_{210}) را از کتابچه کد برمی‌دارد و با استفاده از آن، تصویر را بازسازی می‌کند.

۲. مزیت‌ها نسبت به VAE پایه

مزیت اصلی VQ-VAE حل مشکلی به نام فروپاشی پسین است.

- **مشکل VAE پایه:** در VAEهای معمولی، اگر Decoder خیلی قوی باشد، مدل یاد می‌گیرد که متغیر پنهان z را نادیده بگیرد (یعنی D_{KL} به صفر میل می‌کند) و تصویر را مستقیماً از روی خودش بازسازی کند. این باعث می‌شود فضای پنهان z هیچ چیز معناداری یاد نگیرد.
- **راه‌حل VQ-VAE:**

  - **جلوگیری از فروپاشی:** گسسته‌سازی یک گلوگاه (bottleneck) بسیار شدید ایجاد می‌کند. Decoder مجبور است از اطلاعات فشرده شده در آن اندیس گسسته برای بازسازی تصویر استفاده کند و نمی‌تواند آن را نادیده بگیرد.
  - **فضای پنهان مناسب‌تر:** برای داده‌هایی که ذاتاً گسسته هستند (مانند کلمات یا صدا)، استفاده از فضای پنهان گسسته بسیار طبیعی‌تر و کارآمدتر است.
  - **کیفیت بازسازی:** VQ-VAEها معمولاً تصاویر بسیار واضح‌تری (sharper) نسبت به VAEهای معمولی (که تصاویر تار و محو تولید می‌کنند) بازسازی می‌کنند، چون از آن میانگین‌گیری ناشی از نمونه‌گیری گوسی رنج نمی‌برند.

زیر بخش دوم : VampPrior

۱. توضیح مدل (تخمین توزیع پسین)

VampPrior یک راه‌حل برای مشکل «انعطاف‌ناپذیری» توزیع پیشین (Prior) در VAE استاندارد است.

- **مشکل VAE پایه:** در VAE استاندارد، ما به مدل اجبار می‌کنیم که توزیع پسین تجمعی (aggregated posterior) را در یک توزیع بسیار ساده $N(0, I)$ (گوسی استاندارد) فشرده کند. این توزیع $N(0, I)$ بسیار ساده است و ممکن است نتواند پیچیدگی‌های داده‌های واقعی (مثلاً داده‌های چند-مُدی) را به خوبی مدل کند.
- **راه‌حل VampPrior:** به جای استفاده از $N(0, I)$، ما از یک توزیع پیشین بسیار قوی‌تر و انعطاف‌پذیرتر استفاده می‌کنیم که خود این توزیع پیشین، یک ترکیب از توزیع‌های پسین دیگر است.

۲. نحوه عملکرد

مدل K عدد ورودی ساختگی یاد می‌گیرد. توزیع پیشین $p(z)$ دیگر $N(0, I)$ نیست، بلکه ترکیبی (میانگینی) از K توزیع گوسی است که هر کدام از این توزیع‌ها، خروجی Encoder به ازای یکی از آن ورودی‌های ساختگی

هستند.

$$p(z) = \frac{1}{K}\sum_{i=1}^{K} q(z|x_i)$$

(که x_i ها همان ورودی‌های ساختگیِ در حال یادگیری هستند).

۳. مزیت‌ها

- **پیشین انعطاف‌پذیر:** این توزیع پیشین جدید می‌تواند شکل‌های بسیار پیچیده‌تری (مثلاً چند_قله‌ای) به خود بگیرد که به شکل واقعی توزیع پسین تجمعی داده‌ها نزدیک‌تر است.
- **جلوگیری از KL Vanishing:** چون توزیع پیشین $p(z)$ حالا به اندازه کافی قوی است تا با توزیع پسین $q(z|x)$ مطابقت پیدا کند، ترم D_{KL} دیگر به سادگی به صفر میل نمی‌کند.
- **استفاده بهتر از فضای پنهان:** مدل می‌تواند ساختار غنی‌تری را در فضای پنهان یاد بگیرد و مجبور نیست همه‌ی داده‌ها را در یک گوسی ساده فشرده کند.

زیر بخش سوم : SC-VAE و الگوریتم ISTA

۱. توضیح مدل (SC-VAE)

SC-VAE یک نوع VAE است که هدف آن یادگیری بازنمایی‌های پنهان تنک (Sparse) است.

هدف این است که z از یک توزیعی بیاید که بیشتر درایه‌های بردار z دقیقاً صفر باشند و فقط تعداد کمی از آنها فعال (غیرصفر) باشند، برخلاف VAE استاندارد که z یک بردار چگال (Dense) است.

۲. نقش الگوریتم ISTA

- **مشکل:** Encoderهای استاندارد به خوبی نمی‌توانند خروجی‌های واقعاً تنک تولید کنند.
- **راه‌حل:** به جای یک Encoder استاندارد، این مدل از یک الگوریتم بهینه‌سازی کلاسیک که مخصوص کدگذاری تنک (Sparse Coding) است، استفاده می‌کند.
- ISTA یک الگوریتم تکرارشونده معروف برای حل مسائل بهینه‌سازی تنک (مانند LASSO) است.
- در این مدل، الگوریتم ISTA به صورت باز شده (Unrolled) برای چند گام پیاده‌سازی می‌شود. هر گام از ISTA (که شامل ضرب ماتریسی و یک تابع فعال‌سازی آستانه‌گیری نرم (soft-thresholding)

است) مانند یک لایه در شبکه عصبی عمل می‌کند.

- این شبکه ISTAای باز شده، نقش Encoder را بازی می‌کند. ورودی آن x و خروجی آن، بردار تنک z است. چون کل فرآیند مشتق‌پذیر است، می‌توان از طریق آن backpropagation را انجام داد.

۳. مزیت تنک بودن (Sparsity)

- **جداسازی و تفسیرپذیری:** این بزرگترین مزیت است. یک بازنمایی تنک بسیار قابل تفسیر است.
- **مثال:** اگر z یک بردار ۱۰۰ بُعدی باشد و برای یک تصویر $z = [0, 0, 1.8, 0, ..., -0.9, 0, ..., 0]$ شود، این به وضوح می‌گوید که این تصویر فقط از ویژگی شماره ۳ و ویژگی شماره ۷ تشکیل شده است. در حالی که در VAE استاندارد، z یک بردار چگال است که تفسیر آن تقریباً غیرممکن است.
- **کارایی:** تنک‌سازی یک روش بسیار قوی برای دستیابی به بازنمایی‌های جداشده است.