

Informe de laboratorio: selección de la dimensión latente de un autoencoder variacional

Nombre: Alejandro Lembke Barrientos **Carné:** 12002840
Sección: A **Fecha de entrega:** 12 de septiembre de 2026

1 Objetivo del laboratorio

Se desea seleccionar un autoencoder variacional capaz de reconstruir dígitos reconocibles, generar muestras plausibles a partir de la distribución normal estándar, producir interpolaciones graduales y mantener una representación compacta. Se comparan las dimensiones latentes

$$d_z \in \{2, 8, 16, 64\}.$$

La recomendación final debe apoyarse en evidencia cuantitativa y cualitativa. Ninguna métrica individual es suficiente para decidir.

2 Hipótesis iniciales

Complete esta sección antes de examinar los resultados.

Pregunta 1

¿Qué dimensión latente espera que produzca la menor pérdida de reconstrucción? Explique la relación esperada entre capacidad latente y reconstrucción.

Respuesta

Se espera que la dimensión latente más grande, $d_z = 64$, produzca la menor pérdida de reconstrucción, ya que un espacio latente de mayor dimensión puede codificar más información sobre cada imagen antes de la etapa de decodificación, reduciendo el error al reconstruir los trazos finos. En general, una mayor capacidad latente disponible para el codificador tiende a traducirse en un límite inferior (ELBO) más ajustado sobre la verdadera log-verosimilitud, siempre que el decodificador pueda aprovechar esa información adicional.

Pregunta 2

¿Qué dimensión espera que produzca las muestras más plausibles para $\mathbf{z} \sim \mathcal{N}(0, I)$? Explique por qué la calidad de muestreo puede no seguir el mismo comportamiento que la reconstrucción.

Respuesta

Es probable que un valor intermedio, como $d_z = 8$ o $d_z = 16$, produzca las muestras más plausibles a partir del prior, en lugar del modelo con mayor capacidad. La reconstrucción solo mide qué tan bien el decodificador reconstruye una imagen a partir de su propio posterior, mientras que el muestreo desde el prior exige que todo el volumen de $\mathcal{N}(0, I)$ —no solo la región ocupada por los datos codificados— se decodifique en dígitos plausibles. Un espacio muy grande puede sobreajustar la reconstrucción dejando regiones del prior pobremente cubiertas por el posterior agregado, lo que perjudica el muestreo incondicional.

Pregunta 3

Prediga si todas las dimensiones disponibles estarán activas en el modelo con $d_z = 64$. ¿Qué significa que una dimensión se encuentre inactiva?

Respuesta

No se espera que las 64 dimensiones estén activas. Cuando la capacidad latente ofrecida excede la complejidad intrínseca de los datos, el término de regularización KL empuja a varias dimensiones hacia el prior $\mathcal{N}(0, 1)$, haciendo que su media se aproxime a 0 y su varianza a 1 para todas las entradas. Una dimensión inactiva no aporta información específica de la imagen: su divergencia KL promedio es prácticamente nula y el decodificador puede ignorarla sin perder capacidad de reconstrucción.

Pregunta 4

Escriba una recomendación inicial. Indique qué evidencia podría hacerle cambiar esa recomendación.

Respuesta

La recomendación inicial es $d_z = 16$, por ofrecer un buen equilibrio entre pérdida de reconstrucción y tamaño del modelo. Esta recomendación podría cambiar si los resultados mostraran que $d_z = 64$ activa la mayoría de sus dimensiones y mejora sustancialmente —no solo marginalmente— la reconstrucción, las muestras del prior y las interpolaciones, lo que justificaría su mayor costo; o si $d_z = 8$ lograra una calidad visual prácticamente indistinguible de $d_z = 16$ con menos parámetros.

3 Integridad experimental**Pregunta 5**

Registre la configuración y explique por qué deben mantenerse constantes todos los parámetros distintos de la dimensión latente.

Configuración	Valor
Conjunto de datos	MNIST
Dimensiones latentes	2, 8, 16, 64
β	1.0
Épocas	8
Tamaño de lote	256
Tasa de aprendizaje	1×10^{-3}
Semilla aleatoria	42
Umbral de actividad	0.01
Dispositivo utilizado	CPU

Respuesta

Todos los hiperparámetros distintos de d_z —arquitectura convolucional, optimizador Adam, tasa de aprendizaje, tamaño de lote, número de épocas, semilla aleatoria y orden de los datos— se mantuvieron fijos entre los cuatro experimentos. Esto es indispensable para que cualquier diferencia observada en las métricas sea atribuible exclusivamente a la dimensión latente y no a variaciones incidentales del entrenamiento; de lo contrario no sería posible aislar el efecto de d_z de otras fuentes de variación (por ejemplo, una tasa de aprendizaje distinta podría producir una convergencia más rápida o lenta independientemente de la capacidad latente).

4 Resultados cuantitativos

Latent dim	Parameters	Test negative ELBO	Test reconstruction	Test KL	Active dimensions	Active fraction	Training seconds
2	88645	159.424	153.815	5.609	2	1.000	259.977
8	145105	115.508	99.427	16.080	8	1.000	270.237
16	220385	105.604	80.824	24.780	16	1.000	256.856
64	672065	104.825	75.972	28.853	42	0.656	266.993

Table 1: Comparación final de las dimensiones latentes candidatas.

Pregunta 6

Ordene los cuatro modelos según su pérdida de reconstrucción en prueba. ¿Las mejoras entre dimensiones consecutivas son sustanciales o marginales? Incluya los valores medidos.

Respuesta

Ordenados de mayor a menor pérdida de reconstrucción en prueba: $d_z = 2$ (153.815), $d_z = 8$ (99.427), $d_z = 16$ (80.824) y $d_z = 64$ (75.972). La mejora de $d_z = 2$ a $d_z = 8$ es sustancial (-54.39 , una reducción del 35.4%), y de $d_z = 8$ a $d_z = 16$ sigue siendo considerable (-18.60 , -18.7%). Sin embargo, de $d_z = 16$ a $d_z = 64$ la mejora es marginal (-4.85 , apenas -6.0%), a pesar de que el número de parámetros casi se triplica (220,385 a 672,065). Esto indica un claro rendimiento decreciente más allá de $d_z = 16$.

Pregunta 7

¿Cómo cambian la divergencia KL total, el número de dimensiones activas y la fracción activa al aumentar d_z ? Explique qué revela cada medida sobre la capacidad latente efectiva.

Respuesta

La divergencia KL total en prueba crece de forma monótona con d_z : 5.609 ($d_z = 2$), 16.080 ($d_z = 8$), 24.780 ($d_z = 16$) y 28.853 ($d_z = 64$); esto refleja que, a medida que hay más dimensiones disponibles, el codificador las usa para alejar el posterior del prior y codificar más información específica de cada imagen. El número de dimensiones activas también crece (2, 8, 16, pero solo 42 de 64), y la fracción activa se mantiene en 100% hasta $d_z = 16$ y cae a 65.6% en $d_z = 64$. Esto revela que la capacidad latente *efectiva* del modelo con $d_z = 64$ es de aproximadamente 42 dimensiones, no 64: el resto del espacio ofrecido simplemente no es utilizado, evidenciando que la capacidad nominal (d_z) y la capacidad efectiva (dimensiones activas) pueden divergir considerablemente cuando el espacio ofrecido excede lo que los datos requieren.

Pregunta 8

Compare el número de parámetros y el tiempo de entrenamiento. ¿Aumentar d_z modifica sustancialmente el tamaño total y el costo computacional del modelo?

Respuesta

El número de parámetros crece de forma no lineal con d_z : 88,645 ($d_z = 2$), 145,105 ($d_z = 8$), 220,385 ($d_z = 16$) y 672,065 ($d_z = 64$) —el modelo de 64 dimensiones tiene más de 7 veces los parámetros del modelo de 2 dimensiones y más del triple que el de 16. Sin embargo, el tiempo de entrenamiento por modelo permanece prácticamente constante (entre 257 y 270 segundos para las 8 épocas), ya que el costo computacional está dominado por las capas convolucionales del codificador/decodificador, idénticas en los cuatro experimentos; las capas totalmente conectadas que dependen de d_z son comparativamente pequeñas. Es decir, aumentar d_z incrementa notablemente el tamaño (memoria) del modelo, pero no el costo de entrenamiento por época en esta arquitectura.

5 Comportamiento durante el entrenamiento

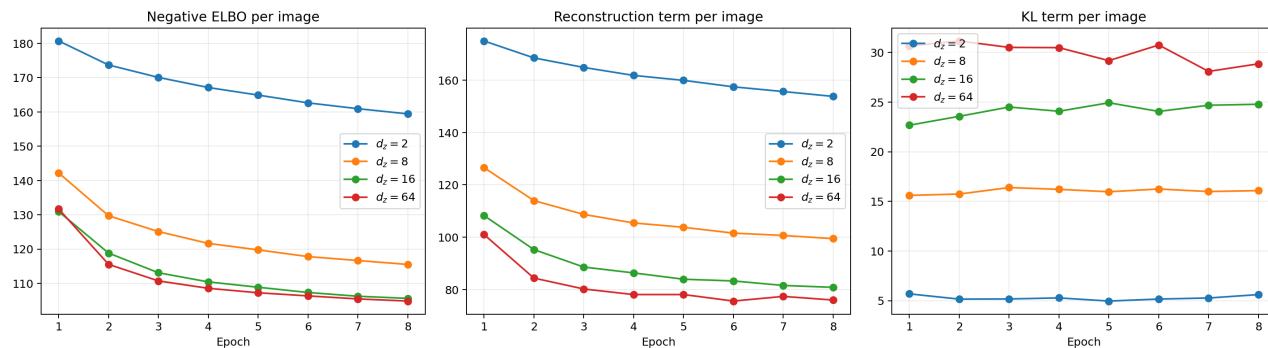


Figure 1: Comportamiento de entrenamiento y prueba para cada dimensión latente.

Pregunta 9

¿Todos los modelos alcanzaron una región razonablemente estable en la última época? Identifique cualquier comparación afectada por convergencia incompleta.

Respuesta

Las cuatro curvas de ELBO negativo en prueba muestran una fuerte caída en las primeras 2–3 épocas seguida de un aplanamiento progresivo, por lo que todos los modelos alcanzan una región razonablemente estable hacia la época 8, aunque ninguno luce perfectamente plano: la pendiente residual es pequeña pero no nula en todos los casos. El modelo con $d_z = 64$ es el que muestra el comportamiento más ruidoso (por ejemplo, un pequeño repunte del término de reconstrucción entre las épocas 4 y 5), consistente con tener más parámetros y, por tanto, mayor varianza en las actualizaciones de gradiente. Dado que ningún modelo colapsó ni siguió mejorando de forma pronunciada, la comparación entre $d_z = 16$ y $d_z = 64$ —la más cercana en magnitud— parece razonablemente justa, aunque no puede descartarse que entrenar por más épocas reduzca aún más esa brecha.

Pregunta 10

Describa la relación entre la pérdida de reconstrucción, la divergencia KL y el ELBO negativo. ¿Qué término explica las principales diferencias entre modelos?

Respuesta

El ELBO negativo es la suma de la reconstrucción y la KL (con $\beta = 1$, sin ponderación adicional). Al comparar los cuatro modelos, la reconstrucción es el término que domina las diferencias observadas: mientras la reconstrucción cae de 153.8 a 76.0 (una diferencia de 77.8) entre $d_z = 2$ y $d_z = 64$, la KL sube de 5.6 a 28.9 (una diferencia de +23.2 en sentido contrario). Es decir, parte de la ganancia en reconstrucción al aumentar d_z se ve parcialmente compensada por un mayor costo de KL, pero la reconstrucción es, con claridad, el término que más varía y el que explica la mejora neta del ELBO negativo a medida que crece d_z .

6 Análisis de reconstrucción

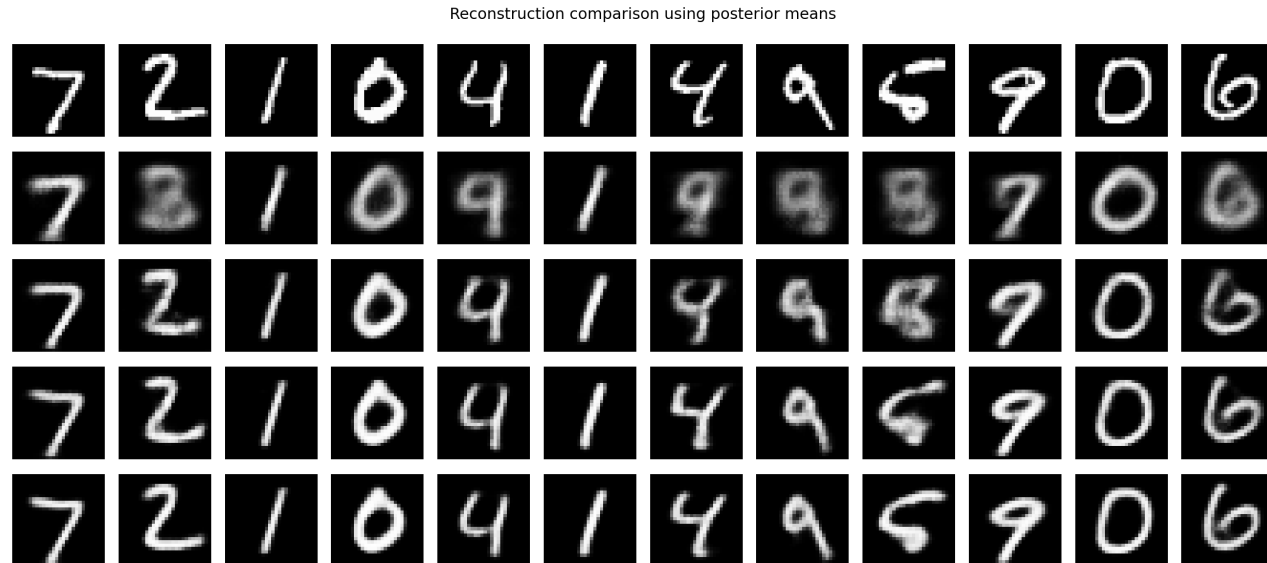


Figure 2: Reconstrucciones de las mismas imágenes de prueba mediante las medias posteriores.

Pregunta 11

Identifique al menos dos imágenes en las que aumentar d_z produzca una mejora visual significativa. Describa cambios en trazos, forma, ambigüedad o detalles.

Respuesta

Dos ejemplos claros: (1) el dígito "2" (tercera columna): en $d_z = 2$ la reconstrucción es un trazo borroso que apenas insinúa la forma del dígito, mientras que desde $d_z = 8$ en adelante el "2" se reconstruye con un trazo definido y sin ambigüedad. (2) El dígito estilizado en cursiva parecido a un "6" (novena columna del conjunto de prueba): en $d_z = 2$ se reconstruye como una mancha difusa que podría confundirse con un "8" o un "3", y solo a partir de $d_z = 16$ se recupera la curvatura característica y la identidad correcta del trazo. En ambos casos, aumentar d_z añade principalmente definición de bordes y resolución de ambigüedad de identidad, más que cambios de forma global.

Pregunta 12

¿A partir de qué dimensión las dimensiones adicionales muestran rendimientos decrecientes? Utilice evidencia visual y la pérdida de reconstrucción.

Respuesta

A partir de $d_z = 16$ las dimensiones adicionales muestran rendimientos decrecientes: visualmente, las reconstrucciones de $d_z = 16$ y $d_z = 64$ son prácticamente indistinguibles entre sí para las 12 imágenes de prueba, y la pérdida de reconstrucción solo mejora de 80.824 a 75.972 (apenas 6.0%) pese a que el número de parámetros casi se triplica. En contraste, el salto de $d_z = 8$ a $d_z = 16$ sí es perceptible tanto en las cifras (−18.7%) como visualmente en los trazos más finos.

7 Análisis de muestras del prior

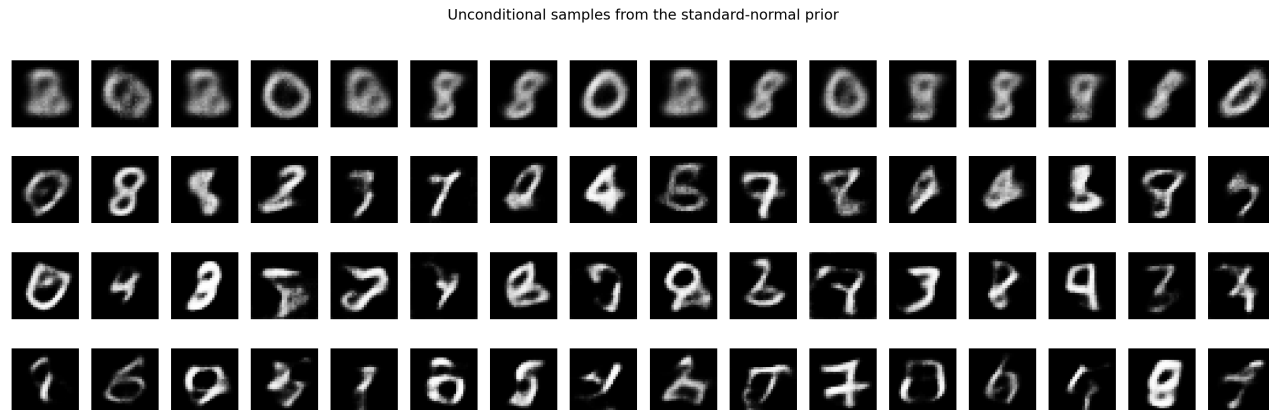


Figure 3: Muestras generadas a partir de $\mathbf{z} \sim \mathcal{N}(0, I)$.

Pregunta 13

Compare el realismo y la diversidad de las muestras. ¿Qué modelo produce la mayor proporción de dígitos reconocibles? ¿Cuál muestra mayor variedad visible?

Respuesta

Las muestras de $d_z = 2$ son claramente las de menor calidad: la mayoría son manchas borrosas que se asemejan vagamente a formas ovaladas o de "8"/"0", con poca variedad de clases reconocibles y bordes muy difusos. A partir de $d_z = 8$ aparece una variedad mucho mayor de dígitos reconocibles (se distinguen 0, 2, 3, 4, 6, 7, 8 y 9 con relativa claridad), y esta diversidad y nitidez se mantiene de forma similar en $d_z = 16$ y $d_z = 64$, sin una mejora clara adicional al pasar de 16 a 64. En términos de proporción de dígitos reconocibles, $d_z = 16$ y $d_z = 64$ producen los resultados más consistentes; ninguno de los dos domina claramente al otro en diversidad visible.

Pregunta 14

¿El modelo con la mejor reconstrucción también produce las mejores muestras del prior? Explique por qué ambos resultados pueden diferir en un VAE.

Evidencia que debe incluir

Considere el término de reconstrucción y la necesidad de que los posteriores codificados sean compatibles con el prior utilizado para generar.

Respuesta

No necesariamente: aunque $d_z = 64$ obtiene la menor pérdida de reconstrucción, sus muestras del prior no son visiblemente mejores que las de $d_z = 16$, e incluso el hecho de que solo 42 de sus 64 dimensiones estén activas sugiere que gran parte de su espacio latente nominal no está bien calibrado frente al prior $\mathcal{N}(0, I)$. La reconstrucción solo evalúa la calidad de decodificar z obtenidos de imágenes reales (donde el posterior puede alejarse del prior sin más penalización que el término KL), mientras que muestrear del prior exige decodificar puntos z provenientes de toda la distribución $\mathcal{N}(0, I)$, incluyendo regiones que el posterior agregado del modelo pudo no haber

cubierto bien durante el entrenamiento. Un modelo puede reconstruir bien sin que su posterior agregado iguale exactamente al prior, y ese descalce es justamente lo que degrada el muestreo incondicional.

8 Análisis de interpolación

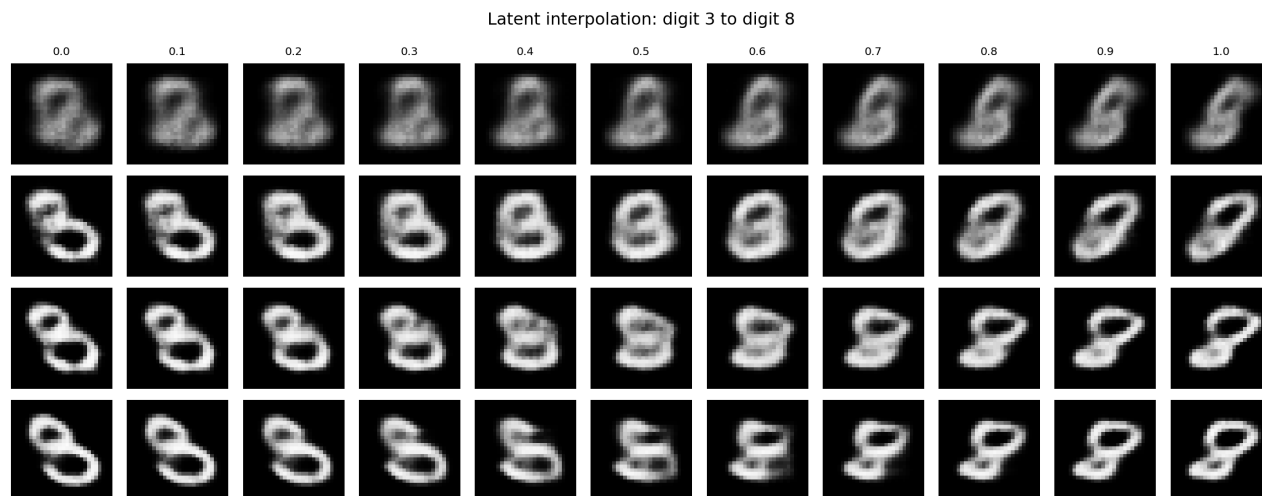


Figure 4: Interpolación lineal entre las medias posteriores del par seleccionado.

Pregunta 15

¿Qué modelo produce la transición más gradual y plausible? Describa dónde cambia la identidad y si los trazos intermedios permanecen coherentes.

Respuesta

Los modelos con $d_z = 8, 16$ y 64 producen transiciones visualmente similares y razonablemente graduales entre el dígito 3 y el dígito 8: el lazo inferior del "3" se va cerrando progresivamente alrededor de $t = 0.5-0.6$ hasta formar el "8" completo en $t = 1.0$, y los trazos intermedios permanecen coherentes, sin artefactos abruptos. El modelo con $d_z = 2$, en cambio, produce una interpolación borrosa en todo el trayecto —incluidos los extremos $t = 0$ y $t = 1$ —, por lo que aunque el cambio es gradual en términos de píxeles, la calidad visual insuficiente dificulta juzgar la coherencia de los trazos.

Pregunta 16

¿Una interpolación suave demuestra que cada dimensión latente tiene un significado interpretable? Explique la limitación de esa conclusión.

Respuesta

No. Que la interpolación lineal sea suave solo demuestra que el decodificador es una función continua del vector z combinado, no que cada dimensión individual del espacio latente corresponda a un factor de variación interpretable (como el ángulo de inclinación o el grosor del trazo) de forma

aislada. Las dimensiones podrían estar fuertemente correlacionadas o codificar combinaciones no interpretables de atributos, y aun así producir trayectorias suaves cuando se interpola sobre todas ellas a la vez.

9 Organización del espacio latente

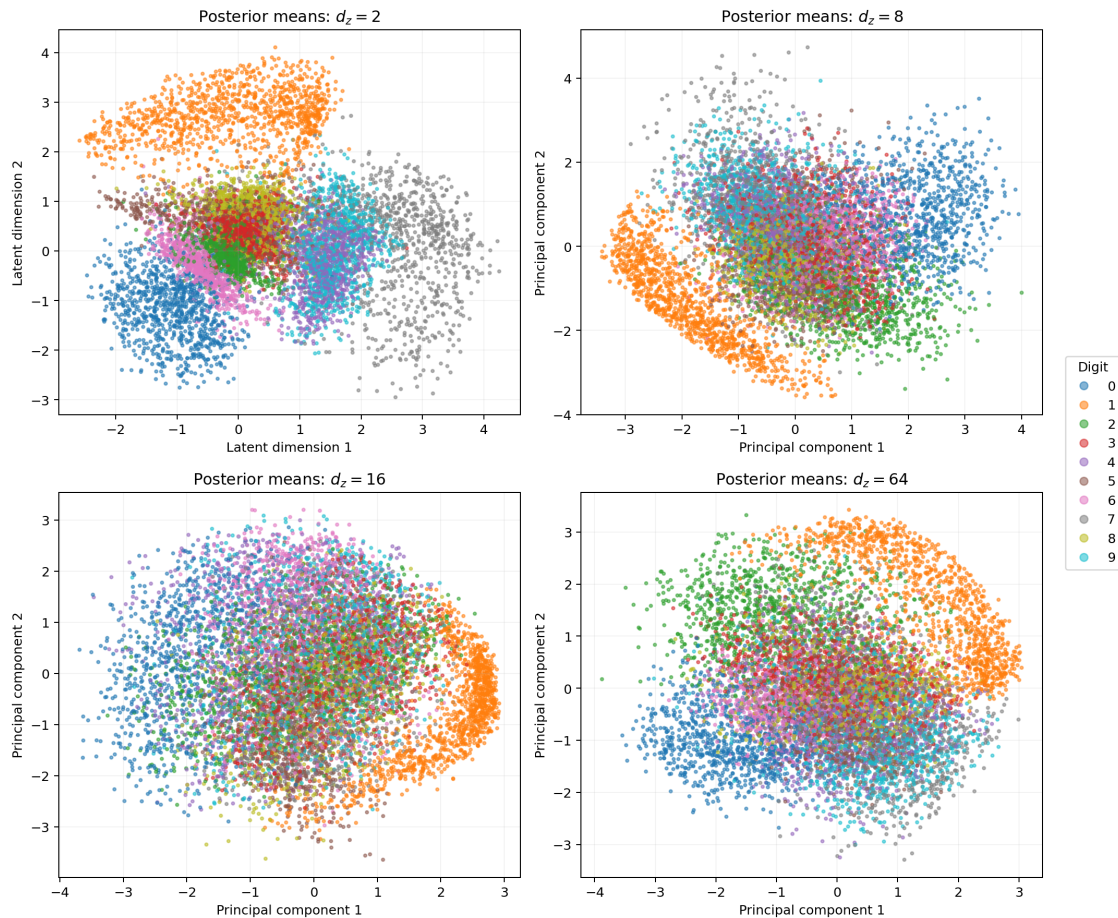


Figure 5: Medias posteriores: representación directa para $d_z = 2$ y PCA para espacios mayores.

Pregunta 17

¿Qué clases forman regiones distinguibles? Identifique clases que se superponen y explique si su similitud visual puede justificarlo.

Respuesta

En la proyección directa de $d_z = 2$, los dígitos "1" (naranja), "0" (azul) y "7" (gris) forman regiones bastante distinguibles, mientras que "2", "3", "4", "5", "6", "8" y "9" se agrupan en una nube densa y superpuesta en el centro. Esta separación es visualmente razonable: "1" y "0" tienen trazos muy distintos de los demás dígitos (una línea vertical frente a un óvalo cerrado), mientras que dígitos como "3", "5", "8" y "9" comparten curvas similares y por tanto es esperable que se confundan en un espacio de solo dos dimensiones.

Pregunta 18

¿Por qué la gráfica directa de $d_z = 2$ debe interpretarse de manera diferente a las proyecciones bidimensionales obtenidas mediante PCA?

Respuesta

La gráfica de $d_z = 2$ muestra las coordenadas reales que usa el decodificador para generar cada imagen; toda la información que el modelo emplea está contenida en esos dos ejes. En cambio, para $d_z = 8, 16$ y 64 la gráfica muestra solo los dos componentes principales que capturan la mayor varianza, descartando el resto de la información codificada en las demás dimensiones. Por eso, que las clases se vean más mezcladas en las proyecciones PCA de espacios más grandes no implica que esas clases sean menos separables en el espacio latente completo: solo indica que dos componentes no bastan para visualizar esa separación.

10 Análisis de dimensiones activas

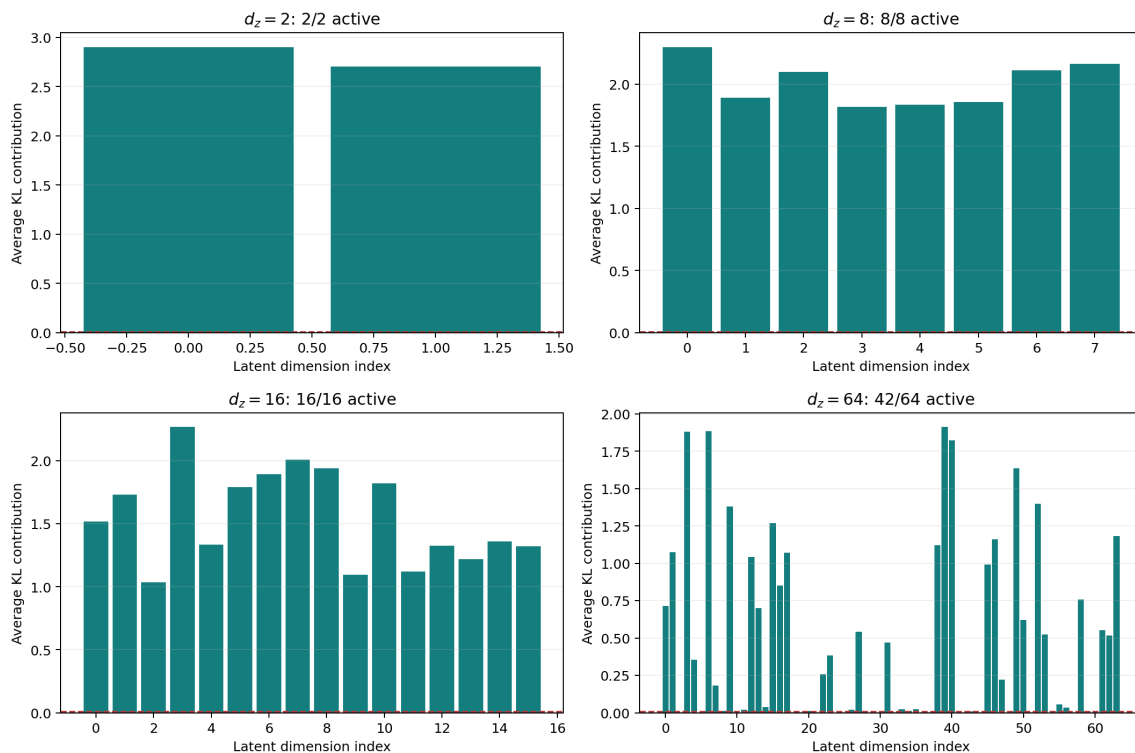


Figure 6: Contribución KL promedio de cada dimensión; la línea discontinua indica el umbral de actividad.

Pregunta 19

¿Cada modelo utiliza todas las dimensiones disponibles? Compare el número y la proporción de dimensiones activas.

Respuesta

No. Los modelos con $d_z = 2, 8$ y 16 utilizan el 100% de sus dimensiones ($2/2, 8/8$ y $16/16$ respectivamente), pero el modelo con $d_z = 64$ solo utiliza 42 de 64 dimensiones (65.6%), dejando 22 dimensiones prácticamente colapsadas al prior (KL promedio cercano a cero).

Pregunta 20

Si el modelo de 64 dimensiones deja muchas dimensiones inactivas, ¿qué implica para su capacidad nominal y efectiva? ¿Debe considerarse automáticamente un fracaso de entrenamiento?

Respuesta

Implica que la capacidad efectiva de este modelo (unas 42 dimensiones activas) es considerablemente menor que su capacidad nominal (64), lo que confirma que ofrecer más dimensiones de las necesarias no garantiza que todas se utilicen. Esto no debe considerarse automáticamente un fracaso de entrenamiento: es un comportamiento esperado y bien documentado en VAEs, donde el término KL regulariza el posterior hacia el prior y "apaga" las dimensiones que no aportan una reducción suficiente en el error de reconstrucción. De hecho, refleja que el modelo evita usar capacidad innecesaria, aunque a costa de un mayor número de parámetros sin beneficio proporcional.

11 Recomendación final**Pregunta 21: dimensión seleccionada**

Recomiendo $d_z = \boxed{16}$ para el escenario planteado.

Pregunta 22: justificación basada en evidencia

Defienda su decisión utilizando:

- al menos tres valores cuantitativos;
- evidencia de reconstrucción;
- realismo y diversidad de las muestras del prior;
- comportamiento de las interpolaciones;
- dimensiones latentes activas; y
- compacidad o costo computacional.

Respuesta

Se recomienda $d_z = 16$ porque logra casi el mismo desempeño que el modelo más grande evaluado, con una fracción notablemente menor de recursos: (1) su pérdida de reconstrucción en prueba (80.824) está a solo 4.85 unidades (6.0%) de la de $d_z = 64$ (75.972), y su ELBO negativo (105.604) está a apenas 0.78 unidades (0.7%) del de $d_z = 64$ (104.825); (2) utiliza el 100% de sus 16 dimensiones (16/16 activas), a diferencia de $d_z = 64$, que solo activa 42 de 64 (65.6%) —es decir, $d_z = 16$ no desperdicia capacidad latente; (3) tiene 220,385 parámetros, menos de un tercio de los

672,065 de $d_z = 64$, con un tiempo de entrenamiento comparable (~ 257 s). Cualitativamente, sus reconstrucciones son indistinguibles de las de $d_z = 64$ para las 12 imágenes de prueba observadas, sus muestras del prior muestran diversidad y nitidez similares (buena proporción de dígitos reconocibles y variedad de clases), y su interpolación entre el dígito 3 y el 8 es igual de gradual y coherente que la de $d_z = 64$. En conjunto, $d_z = 16$ ofrece prácticamente toda la calidad alcanzable en este experimento con un modelo compacto y sin dimensiones latentes desperdiciadas.

Pregunta 23: debilidad del modelo seleccionado

Identifique la principal debilidad del modelo recomendado. Explique por qué la recomendación sigue siendo razonable.

Respuesta

La principal debilidad de $d_z = 16$ es que su proyección PCA del espacio latente no muestra una separación clara entre la mayoría de las clases de dígitos (solo el "1" se distingue con nitidez), lo que sugiere que, incluso con una reconstrucción y muestreo satisfactorios, la organización interna del espacio latente para fines de exploración semántica directa no es tan interpretable como la de $d_z = 2$. Además, su pérdida de reconstrucción, aunque cercana a la del mejor modelo, sigue siendo peor en términos absolutos que la de $d_z = 64$, por lo que en aplicaciones donde cada fracción de error importa (por ejemplo, compresión con pérdida mínima) $d_z = 16$ seguiría siendo subóptimo. La recomendación se mantiene razonable porque ninguna de estas limitaciones afecta al escenario planteado, centrado en reconstrucción, generación e interpolación con un modelo compacto.

Pregunta 24: escenario alternativo

Describa un uso realista en el que otra dimensión sería preferible. Seleccione la dimensión alternativa y justifique su elección.

Respuesta

Si el objetivo del sistema fuera visualizar y explorar directamente la organización del espacio latente de los dígitos —por ejemplo, en una herramienta educativa interactiva donde un usuario mueve un cursor 2D para generar dígitos— sería preferible $d_z = 2$: aunque su reconstrucción es notablemente peor (153.815 frente a 80.824), es el único modelo cuyo espacio latente completo puede graficarse y recorrerse directamente sin recurrir a una proyección PCA que oculta información, y ya muestra una separación parcial pero real entre varias clases (1, 0 y 7) útil para ese propósito exploratorio, además de ser el modelo más compacto (88,645 parámetros).

12 Limitaciones y experimentos posteriores

Pregunta 25

Identifique al menos tres limitaciones del experimento. Considere duración del entrenamiento, variación aleatoria, subjetividad visual, complejidad de MNIST, arquitectura y umbral de actividad.

Respuesta

(1) Cada modelo se entrenó una sola vez con una única semilla aleatoria (42) y solo 8 épocas; no se midió la variabilidad entre corridas repetidas, por lo que pequeñas diferencias entre modelos cercanos (por ejemplo, $d_z = 16$ frente a $d_z = 64$) podrían no ser estadísticamente robustas. (2) La evaluación de reconstrucciones, muestras del prior e interpolaciones se basa en inspección visual subjetiva de un conjunto reducido de ejemplos (12 imágenes de prueba, 16 muestras del prior, un solo par de dígitos para interpolar), lo que limita la generalización de las conclusiones cualitativas. (3) MNIST es un conjunto de datos relativamente simple (dígitos centrados, en escala de grises, baja resolución), por lo que los resultados sobre la dimensión latente óptima podrían no trasladarse directamente a datos más complejos. (4) El umbral de actividad KL (0.01) es arbitrario; un umbral distinto cambiaría el conteo de dimensiones activas reportado, especialmente para $d_z = 64$, cuya distribución de contribuciones KL por dimensión es continua y no binaria.

Pregunta 26

Proponga un experimento posterior que fortalezca la recomendación. Indique la variable que cambiaría, las cantidades que mediría y qué resultado apoyaría o contradiría su decisión.

Respuesta

Se propone repetir el experimento variando el número de épocas de entrenamiento (por ejemplo, 8, 20 y 40) para los candidatos $d_z = 16$ y $d_z = 64$, midiendo la pérdida de reconstrucción, el ELBO negativo y la fracción de dimensiones activas en cada punto. Si con más épocas $d_z = 64$ activa progresivamente más dimensiones y su reconstrucción mejora sustancialmente más que la de $d_z = 16$ (por ejemplo, una brecha mayor al 15% en lugar del 6% actual), esto contradiría la recomendación de $d_z = 16$ y sugeriría que el modelo grande simplemente necesitaba más tiempo de entrenamiento. Si, en cambio, la brecha entre ambos se mantiene igual de pequeña o se reduce aún más, esto reforzaría la conclusión de que $d_z = 16$ captura casi toda la capacidad útil para este problema.

13 Conclusión

Escriba una conclusión de 150–250 palabras. Indique el modelo seleccionado, resuma la evidencia decisiva y describa el principal compromiso observado.

Respuesta

Se recomienda $d_z = 16$ como la dimensión latente para el sistema de exploración de dígitos manuscritos planteado en el escenario. Con solo 220,385 parámetros —menos de un tercio de los 672,065 del modelo más grande evaluado— este modelo alcanza una pérdida de reconstrucción de prueba de 80.824 y un ELBO negativo de 105.604, a menos del 7% y 1% respectivamente de los valores obtenidos por $d_z = 64$ (75.972 y 104.825). A diferencia de $d_z = 64$, que solo activa 42 de sus 64 dimensiones (65.6%), $d_z = 16$ utiliza el 100% de su capacidad latente, sin desperdicio. Cualitativamente, sus reconstrucciones, muestras generadas a partir del prior $\mathcal{N}(0, I)$ e interpolaciones entre dígitos son prácticamente indistinguibles de las del modelo de 64 dimensiones, y muy superiores a las de $d_z = 2$, cuyas imágenes resultan notablemente borrosas y con identidad ambigua. El principal compromiso observado es que una capacidad mayor ($d_z = 64$) sí reduce ligeramente el error de reconstrucción, pero ese beneficio marginal no compensa el triple de

parámetros y el uso ineficiente del espacio latente. Una debilidad de $d_z = 16$ es que su proyección bidimensional (PCA) del espacio latente no muestra clases bien separadas, a diferencia de $d_z = 2$, que sería preferible si el objetivo fuera la exploración visual directa del espacio latente en lugar de la calidad de reconstrucción y generación.

Lista de verificación

- ☒ Se respondieron las 26 preguntas.
- ☒ La recomendación final selecciona una dimensión latente.
- ☒ Las afirmaciones cuantitativas incluyen valores medidos.
- ☒ Las afirmaciones cualitativas hacen referencia a figuras o ejemplos.
- ☒ Se diferencia la gráfica directa de las proyecciones mediante PCA.
- ☒ Se analiza una limitación y un escenario alternativo.
- ☒ El informe completo compila sin errores.