Autoencoders: Aprendizaje de Representaciones Latentes

De organizar el espacio a comprimir información

Cesar Garcia

2025

Introducción

- Retomar la idea de **representaciones internas**
- Introducir autoencoders como modelos **no supervisados**
- Entender el rol del **cuello de botella**
- Preparar el terreno para VAEs

¿Puede una red aprender algo útil sin etiquetas?

¿Qué es un autoencoder?

Definición

Un autoencoder es una red neuronal que aprende a:

- **comprimir** los datos de entrada
- **reconstruirlos** a partir de una representación interna

Entrada = Salida (idealmente)

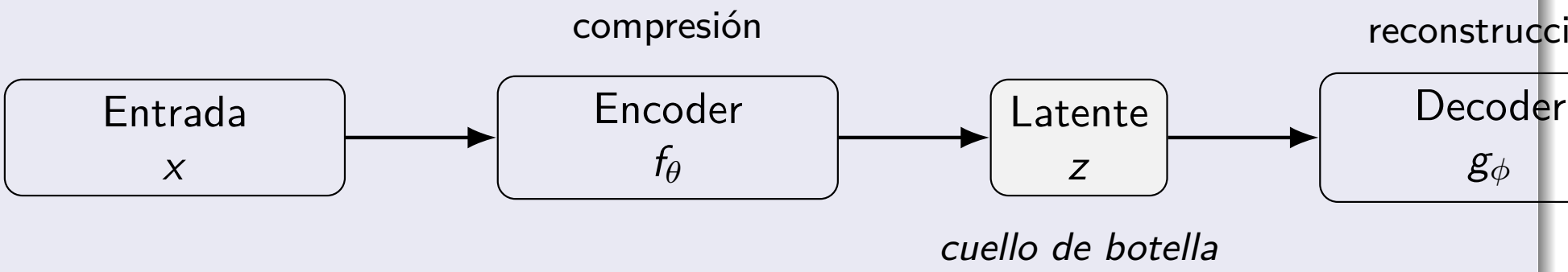
Si la salida es igual a la entrada, ¿qué está aprendiendo realmente?

Arquitectura general

Encoder → Latent Space → Decoder

- **Encoder**: reduce dimensionalidad
- **Latent space**: representación comprimida
- **Decoder**: reconstruye la entrada

La clave no es la reconstrucción, sino **cómo se ve el espacio latente**.



¿Dónde ocurre realmente el aprendizaje importante?

El encoder como extractor

Intuición

El encoder actúa como:

- extractor de características
- organizador del espacio
- función de compresión aprendida

No proyecta al azar: **aprende una geometría útil**.

¿Qué información decide conservar el encoder?

El cuello de botella

Bottleneck

Forzamos al modelo a:

- perder información irrelevante
- conservar estructura esencial

Esto evita la solución trivial de copiar la entrada.

¿Qué pasaría si el espacio latente fuera muy grande?

Función de pérdida

Reconstrucción

La pérdida mide qué tan bien el decoder reconstruye:

- MSE (datos continuos)
- Binary Cross-Entropy (imágenes normalizadas)

No hay etiquetas externas.

¿Qué tipo de errores penaliza esta pérdida?

Autoencoders vs PCA

Comparación conceptual

- PCA: lineal, cerrado, analítico
- Autoencoder: no lineal, flexible, aprendido

Un autoencoder puede verse como:

PCA no lineal entrenado por gradiente

¿Qué ventaja aporta la no linealidad?

Visualizando el espacio latente

Por qué importa

Si el modelo aprendió bien:

- puntos similares $\rightarrow$ cercanos
- clases se organizan espontáneamente

Esto conecta con la sesión anterior.

¿Cómo sabrías si el espacio latente es bueno?

Del encoder al generador (preview)

Limitación clave

Un autoencoder estándar:

- aprende representaciones
- **no es generativo**

No sabemos cómo muestrear su espacio latente.

¿Qué falta para poder generar nuevos datos?

Idea clave de la sesión

Aprender comprimiendo

Un autoencoder:

- no memoriza píxeles
- aprende **qué es esencial**
- organiza el espacio sin supervisión

Representar es decidir qué ignorar.

¿Qué conexión ves con la generalización?