CNN Fundamentals: Convoluciones y Mapas de Características

Por qué funcionan tan bien en imágenes

Cesar Garcia

2025

Introducción

- Problema: una imagen 28×28 tiene **784** entradas
- Una capa densa aprende conexiones con **todos** los píxeles
- En imágenes, la estructura es **local** (bordes, esquinas, texturas)
- Las CNNs explotan esa estructura con **convoluciones**

¿Qué "tipo" de información se repite en distintas partes de una imagen?

¿Por qué no usar solo capas densas?

Intuición

Una red densa para imágenes:

- ignora la vecindad local (un pixel no vive "solo")
- necesita demasiados parámetros
- es sensible a traslaciones (mover el dígito cambia todo)

Las CNNs imponen una hipótesis fuerte:
patrones locales + pesos compartidos

¿Cuál es la ventaja de asumir estructura local?

Convolución: idea central

Filtro que "se desliza"

Una convolución aplica un **mismo filtro** a muchas regiones pequeñas:

- el filtro detecta un patrón (por ejemplo, borde vertical)
- la salida mide "dónde aparece" ese patrón

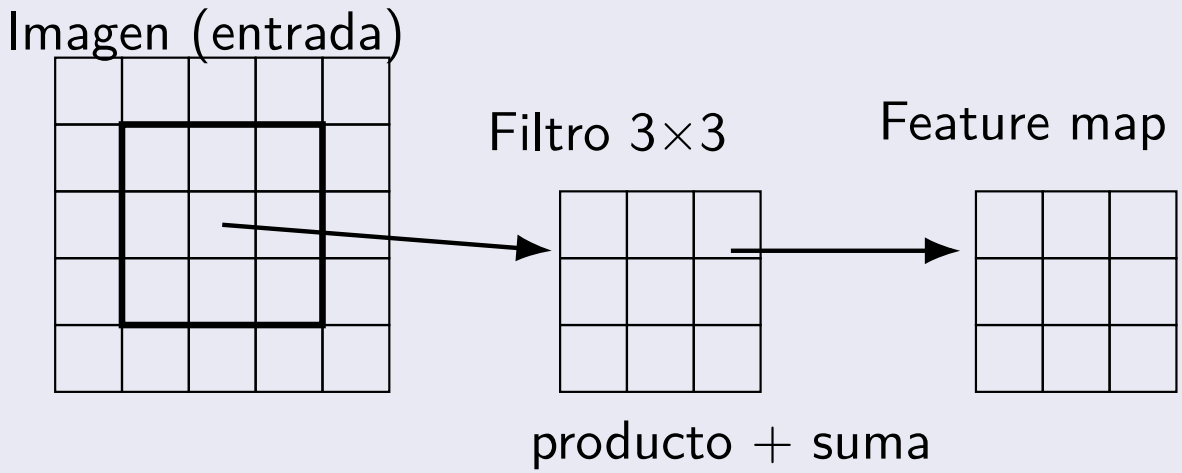
Resultado:

- un **mapa de características** (feature map)

¿Qué significa que el filtro use los "mismos pesos" en toda la imagen?

Convolución (diagrama)

Receptive field local + pesos compartidos



¿Por qué el receptive field pequeño es una buena idea para imágenes?

Canales y filtros

De 1 canal a muchos mapas

Una imagen puede tener:

- 1 canal (grayscale) o 3 (RGB)

Una capa conv con K filtros produce:

- K mapas de características

Cada filtro aprende un patrón diferente.

¿Qué tipo de patrones esperarías en capas conv tempranas?

Stride y Padding

Controlan tamaño y resolución

- **Stride**: cuántos píxeles avanza el filtro
  - mayor stride $\rightarrow$ menos resolución
- **Padding**: ceros alrededor
  - mantiene tamaño y preserva bordes

Regla mental:

stride baja = más detalle

stride alta = más compresión

¿Qué tradeoff estás aceptando al aumentar el stride?

Pooling (resumen)

Reducción y robustez

Pooling (ej. max-pooling) hace:

- reduce resolución
- aumenta invariancia local
- disminuye costo computacional

Pero también:

- puede perder detalle fino

¿Qué tipo de información puede destruir el pooling?

¿Qué aprende una CNN?

Jerarquía de representaciones

Capas conv tempranas:

- bordes, esquinas, texturas

Capas medias:

- partes de objetos

Capas profundas:

- conceptos más abstractos

Esto conecta con Session 8: **representaciones**.

¿Por qué tiene sentido que las representaciones sean jerárquicas?

CNN vs MLP (comparación)

Parámetros y generalización

CNN:

- menos parámetros (pesos compartidos)
- inductive bias útil para imágenes
- mejor generalización con menos datos

MLP:

- puede aprender, pero necesita más datos/parámetros
- no “sabe” que la imagen tiene estructura local

¿Cuándo podría un MLP competir razonablemente con una CNN?

Idea clave de la sesión

Convolución = prior sobre el mundo visual

Las CNNs funcionan porque asumen:

- patrones locales importan
- los mismos patrones pueden aparecer en cualquier lugar
- la jerarquía emerge al apilar capas
  Imponemos estructura para aprender mejor.

¿Qué "supuestos" sobre el mundo están codificados en una CNN?