CNN para visión y retail

De los píxeles a la decisión de negocio: localidad, pesos compartidos, jerarquía.

Conceptos clave

- **Convolución**: un kernel pequeño (típico 3×3) recorre la imagen y produce un *feature map*.
- **Pesos compartidos**: el mismo detector vale en cualquier posición de la imagen. Reduce parámetros varios órdenes de magnitud.
- **Pooling**: reduce resolución, da invariancia a pequeños desplazamientos.
- Arquitecturas canónicas: LeNet → AlexNet → VGG → **ResNet** (skip connections) → EfficientNet → ConvNeXt.

Ejemplo a mano

Convolución 3×3, posición esquina superior izquierda. Parche $P = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \\ 2 & 1 & 0 \end{pmatrix}$, kernel borde vertical $K = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix}$.

$$\text{out}_{0,0} = \sum_{i,j} P_{ij}\, K_{ij} = 1 \cdot 1 + 0 + 2 \cdot (-1) + 0 + 1 \cdot 0 + 3 \cdot (-1) + 2 \cdot 1 + 1 \cdot 0 + 0 = \textbf{-2}.$$

El kernel detecta diferencia izquierda-derecha. Una CNN aprende *muchos* kernels en paralelo.

Concepto → aplicación en retail y operaciones

Modelo	Aplicación	KPI medible
ResNet fine-tuned	Catálogo automático moda (Inditex/Zara).	Top-1, etiquetado/h.
EfficientNet ligera	Control calidad en planta (Mercadona).	Recall defecto, FP/día.
YOLO detección	Conteo en estante (Carrefour).	% rotura stock.
DiT / OCR	Facturas y recibos (BBVA, Endesa).	% revisión humana.
SegFormer U-Net	Imágenes satélite (Iberdrola).	IoU, hectáreas detectadas.

Idea clave del Día 04. El truco no es construir una CNN desde cero: en 2026 cualquier proyecto serio de visión empieza descargando un modelo preentrenado de Hugging Face y entrenando la cabeza. Si tu dataset tiene menos de 50.000 imágenes etiquetadas, **transfer learning es la elección correcta**.

Para llevarte: lectura Prince cap. 10; práctica: fine-tune ResNet-18 sobre Fashion-MNIST o EuroSAT.