

Tesis

Mariana Belén Cruz Rodríguez Yofre Hernán García Gómez

Tabla de contenidos

1	Introducción	4
1.1	Representación de imágenes	4
1.2	Estructura funcional de la red	5
1.3	Parámetros del modelo	5
2	Operación de Convolución	6
2.1	Caso unicanal	6
2.2	Interpretación como producto interno	6
3	Convolución Multicanal	8
3.1	Interpretación tensorial	8
4	Múltiples Filtros (Mapas de Características)	9
5	Interpretación de los kernels	10
5.1	Propiedad de compartición de parámetros	10
6	Resumen	11
7	Stride y Padding	12
7.1	Stride	12
7.2	Padding	12
8	Función de Activación	13
9	Pooling	14
9.1	Max Pooling	14
9.2	Average Pooling	14
10	Capa Totalmente Conectada	15
11	Función de Pérdida	16
11.1	Softmax	16
11.2	Entropía Cruzada	16
12	Entrenamiento	17
13	Propiedades Fundamentales	18
13.1	Conectividad Local	18
13.2	Compartición de Parámetros	18

13.3 Invarianza Traslacional	18
14 Flujo de una CNN	19
15 Interpretación Matemática	20

1 Introducción

Una **red neuronal convolucional** (*Convolutional Neural Network*, CNN) es una arquitectura de aprendizaje profundo diseñada para procesar datos que poseen una estructura de tipo rejilla (es decir, un dominio discreto regular donde los datos están indexados por coordenadas enteras y presentan relaciones locales bien definidas), como es el caso de las imágenes digitales.

Desde un punto de vista formal, una CNN puede entenderse como una función parametrizada

$$f(\cdot, \theta) : \mathcal{X} \subseteq \mathbb{R}^{H \times W \times C} \longrightarrow \mathbb{R}^k,$$

donde:

- $\mathcal{X}$ es el espacio de entrada (por ejemplo, imágenes),
- H representa la **altura** (*height*) de la imagen, es decir, el número de filas,
- W representa el **ancho** (*width*), es decir, el número de columnas,
- C representa el **número de canales** (*channels*), que corresponde a la cantidad de valores asociados a cada píxel,
- k es la dimensión del espacio de salida (por ejemplo, número de clases en clasificación),
- θ denota el conjunto total de parámetros entrenables del modelo.

1.1. Representación de imágenes

Una imagen se modela matemáticamente como un tensor de orden 3:

$$X \in \mathbb{R}^{H \times W \times C}.$$

Esto significa que:

- Para cada posición espacial (i, j) con $1 \leq i \leq H$, $1 \leq j \leq W$,
- el valor $X(i, j)$ no es escalar, sino un vector en $\mathbb{R}^C$:

$$X(i, j) = (x_1, x_2, \dots, x_C).$$

1.2. Estructura funcional de la red

Una CNN se construye como una composición de funciones:

$$f(x; \theta) = f^{(L)} \circ f^{(L-1)} \circ \dots \circ f^{(1)}(x),$$

donde:

- $L \in \mathbb{N}$ es el número total de capas de la red,
 - cada $f^{(l)}$ es una función asociada a la capa l -ésima.
-

1.3. Parámetros del modelo

El conjunto de parámetros θ contiene todos los elementos ajustables del modelo:

$$\theta = \theta^{(1)}, \theta^{(2)}, \dots, \theta^{(L)}.$$

En una CNN, estos parámetros incluyen principalmente:

- **kernels** (filtros convolucionales),
- **sesgos** (bias) asociados a cada filtro.

Estos parámetros determinan completamente la forma de la función $f(x; \theta)$ y son aprendidos a partir de los datos mediante optimización.

2 Operación de Convolución

La operación fundamental en una CNN es la **convolución discreta** (en la práctica, correlación cruzada).

2.1. Caso unicanal

Sea una imagen:

$$X \in \mathbb{R}^{H \times W}$$

y un kernel (o filtro):

$$K \in \mathbb{R}^{k_h \times k_w},$$

donde:

- k_h : altura del kernel,
- k_w : ancho del kernel.

La salida Y se define como:

$$Y(i, j) = \sum_{u=0}^{k_h-1} \sum_{v=0}^{k_w-1} K(u, v) X(i+u, j+v).$$

2.2. Interpretación como producto interno

Sea $X_{(i,j)}$ la submatriz de X de tamaño $k_h \times k_w$:

$$X_{(i,j)} = X(i+u, j+v)_{u,v}.$$

Entonces:

$$Y(i, j) = \langle K, X_{(i,j)} \rangle,$$

donde $\langle \cdot, \cdot \rangle$ denota el producto interno en $\mathbb{R}^{k_h \times k_w}$.

Esto implica que la convolución mide la **similitud entre el kernel y una región local de la imagen**.

3 Convolución Multicanal

En el caso general, la entrada es:

$$X \in \mathbb{R}^{H \times W \times C},$$

y el kernel:

$$K \in \mathbb{R}^{k_h \times k_w \times C}.$$

La convolución se define como:

$$Y(i, j) = \sum_{c=1}^C \sum_{u=0}^{k_h-1} \sum_{v=0}^{k_w-1} K(u, v, c), X(i+u, j+v, c).$$

3.1. Interpretación tensorial

Si se vectoriza cada ventana:

$$X_{(i,j)} \in \mathbb{R}^{k_h k_w C}, \quad K \in \mathbb{R}^{k_h k_w C},$$

entonces:

$$Y(i, j) = K^\top X_{(i,j)}.$$

Esto muestra que la convolución es un operador lineal local.

4 Múltiples Filtros (Mapas de Características)

En una CNN no se utiliza un solo kernel, sino un conjunto de filtros:

$$K^{(f)} \in \mathbb{R}^{k_h \times k_w \times C}, \quad f = 1, \dots, F.$$

Cada filtro produce un mapa de activación distinto:

$$Y_f(i, j) = \sum_{c=1}^C \sum_{u=0}^{k_h-1} \sum_{v=0}^{k_w-1} K^{(f)}(u, v, c), X(i+u, j+v, c).$$

Por tanto, la salida total es un tensor:

$$Y \in \mathbb{R}^{H' \times W' \times F},$$

donde:

- H' y W' dependen del tamaño del kernel, el padding y el stride,
 - F es el número de filtros, es decir, el número de **canales de salida**.
-

5 Interpretación de los kernels

Cada kernel puede interpretarse como un **detector de patrones locales**.

Formalmente:

$$Y_f(i, j) = \langle K^{(f)}, X_{(i,j)} \rangle.$$

Por tanto:

- Si $Y_f(i, j)$ es grande, entonces la región $X_{(i,j)}$ es similar al patrón definido por $K^{(f)}$.
 - Si es pequeño o negativo, no hay coincidencia significativa.
-

5.1. Propiedad de compartición de parámetros

Un aspecto fundamental es que el mismo kernel $K^{(f)}$ se aplica en todas las posiciones (i, j) :

$$K^{(f)}(u, v, c) \text{ es independiente de } (i, j).$$

Esto implica:

- reducción significativa del número de parámetros,
 - capacidad de detectar patrones independientemente de su posición (invarianza traslacional).
-

6 Resumen

Una CNN puede entenderse como una función parametrizada

$$f(x; \theta) = f^{(L)} \circ \dots \circ f^{(1)}(x),$$

donde:

- $x \in \mathbb{R}^{H \times W \times C}$ es una imagen,
 - cada capa aplica operadores lineales locales (convoluciones) seguidos de no linealidades,
 - los parámetros θ están formados principalmente por kernels,
 - cada kernel actúa como un detector de patrones mediante productos internos locales.
-

7 Stride y Padding

7.1. Stride

El stride s controla el desplazamiento del kernel. Las dimensiones de salida son:

$$H' = \left\lfloor \frac{H - k_h + 2p}{s} \right\rfloor + 1, \quad W' = \left\lfloor \frac{W - k_w + 2p}{s} \right\rfloor + 1.$$

7.2. Padding

Se define como la adición de ceros alrededor de la imagen:

$$X_{\text{pad}} \in \mathbb{R}^{(H+2p) \times (W+2p) \times C}.$$

8 Función de Activación

Después de la convolución se aplica una función no lineal. Un ejemplo común es la función ReLU:

$$\phi(x) = \text{máx}(0, x).$$

La salida es:

$$Z = \phi(Y).$$

9 Pooling

El pooling reduce la dimensionalidad.

9.1. Max Pooling

$$Y(i, j) = \max_{(u, v) \in \mathcal{R}} X(i + u, j + v).$$

9.2. Average Pooling

$$Y(i, j) = \frac{1}{|\mathcal{R}|} \sum_{(u, v) \in \mathcal{R}} X(i + u, j + v).$$

10 Capa Totalmente Conectada

Se transforma el tensor en un vector:

$$X \in \mathbb{R}^{H \times W \times C} \rightarrow x \in \mathbb{R}^n.$$

Luego:

$$y = Wx + b,$$

donde $W \in \mathbb{R}^{m \times n}$ y $b \in \mathbb{R}^m$.

11 Función de Pérdida

11.1. Softmax

$$\hat{y}_i = \frac{e^{z_i}}{\sum_j e^{z_j}}.$$

11.2. Entropía Cruzada

$$\mathcal{L} = - \sum_i y_i \log(\hat{y}_i).$$

12 Entrenamiento

El problema de optimización es:

$$\min_{\theta} \mathcal{L}(f(x; \theta), y).$$

Se utiliza descenso de gradiente:

$$\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}.$$

13 Propiedades Fundamentales

13.1. Conectividad Local

Cada neurona depende de una región pequeña de la entrada.

13.2. Compartición de Parámetros

El mismo kernel se aplica en toda la imagen.

13.3. Invarianza Traslacional

El modelo detecta patrones independientemente de su posición.

14 Flujo de una CNN

Imagen $\rightarrow$ Convolución $\rightarrow$ ReLU $\rightarrow$ Pooling $\rightarrow \dots \rightarrow$ Fully Connected $\rightarrow$ Softmax

15 Interpretación Matemática

Una CNN puede verse como una función:

$$f : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^k,$$

compuesta por operadores lineales, no lineales y de reducción.

Las redes neuronales convolucionales son modelos altamente eficientes para el procesamiento de imágenes debido a la combinación de convoluciones, no linealidades y reducción de dimensionalidad, junto con propiedades estructurales como la compartición de parámetros y la conectividad local.