Introducción a CNN y Reconocimiento de Emociones

Diseño de Interfaces y UX

Dr. Hugo Arnoldo Mitre Hernández

CIMAT Zacatecas

Marzo 2026

¿Qué es una imagen?

Visión computacional profunda

- Una imagen es un arreglo de pixeles, la cual puede tener 1 o más canales de color. Usualmente:
  - 1 canal de color $\rightarrow$ Escala de grises
  - 3 canales de color $\rightarrow$ Escala RGB
  - 4 canales de color $\rightarrow$ Escala RGBA
- Un pixel puede ser visto como un objeto 5-dimensional (x, y, r, g, b).

Max-Feature-Map (MFM)

Definición

MFM es una función de activación que selecciona la característica más relevante entre dos mapas de características.

Idea clave:

- Sustituye a ReLU
- Introduce competencia entre neuronas
- Reduce dimensionalidad

Ecuación de MFM

$$MFM(x_1, x_2) = \text{máx}(x_1, x_2)$$

Donde:

- x_1, x_2: mapas de características
- Se comparan elemento a elemento

Resultado:

- Se conserva el valor más alto
- Se descarta la información menos relevante

Ejemplo de MFM

Entrada:

$$x_1 = \begin{bmatrix} 1 & 5 \\ 3 & 2 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 & 2 \\ 1 & 6 \end{bmatrix}$$

Salida (MFM):

$$\text{máx}(x_1, x_2) = \begin{bmatrix} 4 & 5 \\ 3 & 6 \end{bmatrix}$$

Interpretación de MFM

- Cada par de filtros compite entre sí
- Solo la característica más fuerte sobrevive

Razonamiento

MFM actúa como un mecanismo de selección automática de características relevantes.

En rostros:

- Un filtro detecta borde de ojo
- Otro detecta ruido
- MFM conserva el borde

MFM vs ReLU

Propiedad	ReLU	MFM
Operación	máx$(0, x)$	máx(x_1, x_2)
Entrada	1 mapa	2 mapas
Salida	Igual tamaño	Mitad de canales
Selección	Umbral	Competencia

Ventaja de MFM:

- Reduce parámetros
- Mejora generalización

MFM en Light CNN-9

- Cada convolución genera $2N$ mapas
- MFM reduce a N mapas

Ejemplo:

- Conv $\rightarrow$ 96 canales
- MFM $\rightarrow$ 48 canales

Impacto

Reduce el costo computacional y elimina redundancia

MFM en reconocimiento de emociones

- Filtra características irrelevantes
- Mejora la detección de patrones faciales

Ejemplo:

- Sonrisa $\rightarrow$ patrón fuerte
- Ruido $\rightarrow$ patrón débil
- MFM conserva la sonrisa

Resultado:

- Modelo más robusto
- Mejor clasificación emocional

Arquitectura Light CNN-9

Capa	Tipo	Filtros / Unidades	Salida
Input	Imagen	-	$128 \times 128 \times 1$
Conv1	$5 \times 5 + \text{MFM}$	$96 \rightarrow 48$	$128 \times 128 \times 48$
Pool1	MaxPool 2×2	-	$64 \times 64 \times 48$
Conv2a	$1 \times 1 + \text{MFM}$	$96 \rightarrow 48$	$64 \times 64 \times 48$
Conv2	$3 \times 3 + \text{MFM}$	$192 \rightarrow 96$	$64 \times 64 \times 96$
Pool2	MaxPool 2×2	-	$32 \times 32 \times 96$
Conv3a	$1 \times 1 + \text{MFM}$	$192 \rightarrow 96$	$32 \times 32 \times 96$
Conv3	$3 \times 3 + \text{MFM}$	$384 \rightarrow 192$	$32 \times 32 \times 192$
Pool3	MaxPool 2×2	-	$16 \times 16 \times 192$
Conv4a	$1 \times 1 + \text{MFM}$	$384 \rightarrow 192$	$16 \times 16 \times 192$
Conv4	$3 \times 3 + \text{MFM}$	$256 \rightarrow 128$	$16 \times 16 \times 128$
Pool4	MaxPool 2×2	-	$8 \times 8 \times 128$
Conv5a	$1 \times 1 + \text{MFM}$	$256 \rightarrow 128$	$8 \times 8 \times 128$
Conv5	$3 \times 3 + \text{MFM}$	$256 \rightarrow 128$	$8 \times 8 \times 128$
Pool5	MaxPool 2×2	-	$4 \times 4 \times 128$
FC1	Fully Connected + MFM	$512 \rightarrow 256$	256
Output	Clasificación	N clases	N

Arquitectura Light CNN-9 — Lectura general

Objetivo de la red

Tomar una imagen de cara en escala de grises (128×128 px) y producir un vector de identidad de 256 dimensiones, comprimiendo progresivamente el espacio espacial mientras se extraen características discriminativas.

Patrón que se repite 5 veces

Progresión espacial

$$128 \xrightarrow{\text{Pool1}} 64 \xrightarrow{\text{Pool2}} 32 \xrightarrow{\text{Pool3}} 16 \xrightarrow{\text{Pool4}} 8 \xrightarrow{\text{Pool5}} 4$$

Cómo leer la tabla bloque a bloque

1. Entrada
Imagen $128 \times 128 \times 1$: 128 píxeles de ancho, 128 de alto, 1 canal (gris).

2. Conv1 + Pool1
Conv 5×5 produce 96 mapas; MFM los reduce a 48.
Pool2×2 reduce la resolución a la mitad: 64×64.

3. Conv2a + Conv2 + Pool2
Conv 1×1 como *bottleneck* comprime canales antes de la conv 3×3, reduciendo costo computacional. Pool lleva a 32×32.

4. Bloques 3–5
Mismo patrón: bottleneck 1×1 + conv 3×3 + MFM + Pool. La resolución baja: $16 \rightarrow 8 \rightarrow 4$ px. Los canales suben: $96 \rightarrow 192 \rightarrow 128$.

5. FC1 + MFM
Los $4 \times 4 \times 128 = 2{,}048$ valores se aplanan y se conectan a 512 neuronas; MFM las reduce a **256**: el vector de identidad.

6. Salida
Los 256 valores se mapean a N clases según la tarea de clasificación.

Puntos clave para discutir:

- ¿Por qué usar conv 1×1 antes de conv 3×3?
- ¿Qué pasa con la resolución espacial después de cada Pool?
- ¿Por qué el número de canales sube y luego baja?
- ¿En qué se diferencia MFM de ReLU conceptualmente?

Gracias