VGGNet

7 de junio de 2025

Según Simonyan and Zisserman [2015] y Russakovsky et al. [2015], el grupo Visual Geometry Group de la universidad de Oxford desarrolló VGGNet, una arquitectura de redes neuronales convolucionales. Su objetivo principal era estudiar cómo afecta la profundidad a la precisión en tareas de clasificación y reconocimiento de imágenes. Las versiones más conocidas fueron VGG-16 y VGG-19, con 16 y 19 capas convolucionales respectivamente. En la competición ILSVRC-2014, la versión VGG-16 alcanzó un top-5 accuracy del 92,7 % sobre el conjunto de datos de ImageNet, el cual incluye más de 14 millones de imágenes divididas en 1000 categorías.

Su arquitectura se basa en una estructura de red profunda que utiliza filtros de convolución del mínimo tamaño en todas sus capas. El término "profundo" se refiere al gran número de capas que tiene, lo que permite a la red aprender representaciones más abstractas a medida que se avanza hacia capas más profundas. VGG-16 tiene 13 capas convolucionales, mientras que VGG-19 tiene 16 y ambos modelos están seguidos de tres capas totalmente conectadas.

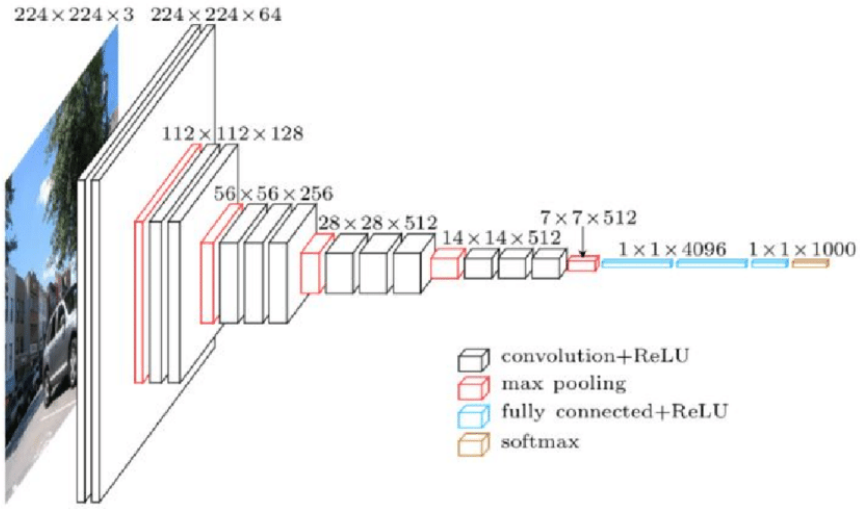

Figura 1: Estructura de capas de VGGNet tomada de Alonso [2020]

VGGNet, como parte de su diseño para la competición ILSVRC, acepta imágenes de entrada con una resolución estándar de 224x224 píxeles. Este tamaño fue elegido para asegurar la consistencia en el tamaño de entrada durante el proceso de entrenamiento y evaluación del modelo en el conjunto de datos ImageNet.

En las capas convolucionales se utilizan filtros de tamaño mínimo, es decir de 3x3. Esto ayuda a reducir el coste computacional, en comparación a filtros más grandes, mientras se siguen capturando características complejas de la imagen. Estas capas están seguidas de activaciones ReLU que ayudan a introducir no linealidad, mejorar el proceso de aprendizaje y acelerar el entrenamiento. Al final de cada bloque de capas de convolución se aplica una operación de *max-pooling* que reduce la dimensionalidad de las imágenes preservando las características más importantes. Esto ayuda a reducir el coste computacional y evitar el sobreajuste.

Finalmente se encuentran tres capas totalmente conectadas. La primera contiene 25088 neuronas seguida de dos capas de 4096 neuronas cada una. Por último hay otra capa más que corresponde a la salida con 1000 neuronas que representan cada una de las categorías del conjunto de datos ImageNet.

Layer	**Feature Map**	**Size**	**Kernel Size**	**Stride**	**Activation**
Input	Image	$224 \times 224 \times 3$	-	-	-
1	2 X Convolution	$224 \times 224 \times 64$	3×3	1	ReLU
	Max Pooling	$112 \times 112 \times 64$	3×3	2	-
3	2 X Convolution	$112 \times 112 \times 128$	3×3	1	ReLU
	Max Pooling	$56 \times 56 \times 128$	3×3	2	-
5	2 X Convolution	$56 \times 56 \times 256$	3×3	1	ReLU
	Max Pooling	$28 \times 28 \times 256$	3×3	2	-
7	3 X Convolution	$28 \times 28 \times 512$	3×3	1	ReLU
	Max Pooling	$14 \times 14 \times 512$	3×3	2	-
10	3 X Convolution	$14 \times 14 \times 512$	3×3	1	ReLU
	Max Pooling	$7 \times 7 \times 512$	3×3	2	-
13	Fully Connected (FC)	25088	-	-	ReLU
14	Fully Connected (FC)	4096	-	-	ReLU
15	Fully Connected (FC)	4096	-	-	ReLU
Output	Fully Connected (FC)	1000	-	-	Softmax

Cuadro 1: Estructura de las capas de la red VGGNet

La arquitectura de VGGNet es relativamente sencilla de implementar y entender. Esta simplicidad, combinada con su efectividad en tareas de clasificación de imágenes, ha hecho que VGGNet se convierta en un referente en el campo de las CNN. Debido a su capacidad para aprender representaciones profundas de imágenes, VGGNet ha sido muy utilizada para transferencia de aprendizaje, donde una red preentrenada en un gran conjunto de datos como ImageNet se ajusta para realizar tareas específicas en conjuntos de datos más pequeños.

Uno de los principales inconvenientes de VGGNet es la cantidad masiva de parámetros que posee. Con más de 138 millones de parámetros, el modelo requiere grandes cantidades de memoria y potencia computacional, lo que ralentiza considerablemente el proceso de entrenamiento y lo hace costoso en términos de recursos. Este gran número de parámetros también hace que la red tienda al sobreajuste.

Referencias

Diego Gabriel Alonso. *Enfoque* liviano *para reconocimiento de gestos manuales híbridos con cámaras de profundidad*. PhD thesis, Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil, Argentina, 2020. Tesis Doctoral, accedido: Diciembre 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. URL https://arxiv.org/abs/1409.0575. Accedido: Febrero 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.1556. Accedido: Febrero 2025.