

Evaluación de Operaciones de Pooling en Arquitecturas convolucionales para objetos Reconocimiento

Dominik Scherer, Andreas Müller, y Sven Behnke

Universidad de Bonn, Instituto de Ciencias de la Computación
VI, Grupo de Sistemas Inteligentes Autónomos, Römerstr. 164, 53117 Bonn, Alemania
{scherer|amueller}@ais.uni-bonn.de, behnke@cs.uni-bonn.de <http://www.ais.uni-bonn.de>

Abstracto. Una práctica común para obtener características invariantes en modelos de reconocimiento de objetos es agregar múltiples características de bajo nivel en un vecindario pequeño. Sin embargo, las diferencias entre esos modelos dificultan la comparación de las propiedades de diferentes funciones de agregación. Nuestro objetivo es obtener información sobre diferentes funciones comparándolas directamente en una arquitectura fija para varias tareas comunes de reconocimiento de objetos. Los resultados empíricos muestran que una operación de agrupación máxima supera significativamente a las operaciones de submuestreo. A pesar de sus propiedades de invariante de desplazamiento, las ventanas de agrupación superpuestas no representan una mejora significativa con respecto a las ventanas de agrupación que no se superponen. Al aplicar este conocimiento, logramos tasas de error de última generación del 4,57 % en el conjunto de datos uniforme normalizado NORB y del 5,6 % en el conjunto de datos desordenados NORB.

1. Introducción

Muchas arquitecturas recientes de reconocimiento de objetos se basan en el modelo de corteza visual de mamíferos propuesto por Hubel y Wiesel [8]. Según sus hallazgos, el área visual V1 se compone de células simples y células complejas. Mientras que las celdas simples realizan una extracción de características, las celdas complejas combinan varias de estas características locales de una pequeña vecindad espacial. Se supone que la agrupación espacial es crucial para obtener características invariantes en la traducción.

Los modelos supervisados basados en esos hallazgos son el Neocognitron [6] y las redes neuronales convolucionales (CNN) [10]. Muchos extractores de características de última generación emplean técnicas de agregación similares, incluidos histogramas de gradientes orientados (HOG) [3], descriptores SIFT [12], características Gist [22] y el modelo HMAX [20].

Estos modelos se pueden distinguir ampliamente por la operación que resume una vecindad espacial. La mayoría de los modelos anteriores realizan una operación de submuestreo, donde el promedio de todos los valores de entrada se propaga a la siguiente capa. Dichas arquitecturas incluyen el Neocognitrón, las CNN y la Pirámide de Abstracción Neural [2]. Un enfoque diferente es calcular el valor máximo en un

Este trabajo fue apoyado en parte por NRW State dentro de la Escuela de Investigación B-IT.

vecindario. Esta dirección la siguen el modelo HMAX y algunas variantes de las CNN.

Si bien se han comparado ampliamente modelos completos, hasta el momento no se ha realizado ninguna investigación que evalúe la elección de la función de agregación. Por tanto, el objetivo de nuestro trabajo es determinar empíricamente cuál de las funciones de agregación establecidas es más adecuada para tareas de visión. Además, investigamos si las ideas del procesamiento de señales, como la superposición de campos receptivos y funciones de ventana, pueden mejorar el rendimiento del reconocimiento.

2. Trabajo relacionado

Muchas arquitecturas de visión por computadora inspiradas en estudios de la corteza visual primaria utilizan un procesamiento en múltiples etapas de células simples y complejas. Las celdas simples realizan la detección de características en alta resolución. La invariancia de traducción y la generalización se logran mediante células complejas, que combinan activaciones en una vecindad local.

Uno de los primeros modelos que empleó esta técnica es el Neocognitron [6]. Aquí, cada una de las llamadas células C recibe conexiones de entrada excitadoras de células de extracción de características en posiciones ligeramente diferentes. Una célula C se vuelve activa si al menos una de sus entradas está activa, tolerando así ligeras deformaciones y transformaciones.

En las redes neuronales convolucionales (CNN), como LeNet-5 [10], la invariancia de cambio se logra con capas de submuestreo. Las neuronas de estas capas reciben información de un pequeño campo receptivo no superpuesto de la capa anterior. Cada neurona calcula la suma de sus entradas, la multiplica por un coeficiente entrenable, agrega un sesgo entrenable y pasa el resultado a través de una función de transferencia no lineal. Se realiza un cálculo similar en la Pirámide de Abstracción Neural recurrente [2]. Más recientemente, la operación de submuestreo en las CNN ha sido reemplazada por una operación de agrupación máxima [18]. Aquí, sólo el valor máximo dentro del campo receptivo se propaga a la siguiente capa.

En la descripción de escena global calculada por el modelo Gist [22], el extractor de características no se puede entrenar, pero realiza cálculos similares. Las características envolventes centrales de bajo nivel se calculan a partir de canales de color e intensidad en diferentes escalas y orientaciones. Posteriormente, cada canal se divide en una grilla de subregiones de 4×4 . El vector de características Gist de 16 dimensiones de este canal se calcula promediando los valores en cada región.

El descriptor de punto clave SIFT [12] (transformación de característica invariante de escala) se calcula muestreando la orientación dominante y la magnitud del gradiente alrededor de la ubicación del punto clave. Estos valores, ponderados por una función de ventana gaussiana, luego se acumulan en matrices de histogramas de orientación de 4×4 que resumen el contenido en posiciones de 4×4 . Las pirámides de histogramas de orientación local calculadas sobre una cuadrícula densa y superpuesta de parches de imágenes han demostrado ser representaciones de escenas y objetos muy adecuadas para tareas de reconocimiento [9]. Experimentos con histograma de gradientes orientados (HOG) normalizado localmente

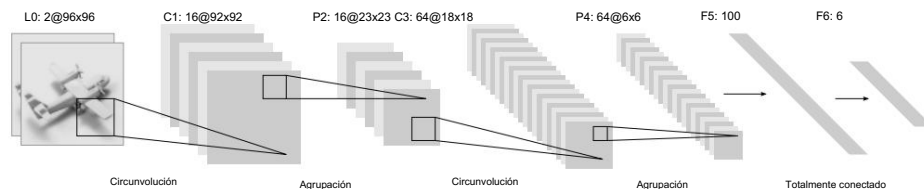


Fig. 1. Arquitectura de nuestra CNN para experimentos NORB, que consta de alternancia capas convolucionales y de agrupación. Las capas de agrupación pueden implementar cualquier submuestreo operaciones o agrupación máxima.

Los descriptores han demostrado que los gradientes de escala fina y la agrupación espacial de escala gruesa producen un buen rendimiento de reconocimiento para la detección humana [3].

Riesenhuber y Poggio propusieron originalmente utilizar una operación máxima para modelar células complejas en 1999 [20]. En el modelo HMAX [21], cada complejo La celda recibe información de un parche distinto de un mapa de celda simple. Este enfoque fue perfeccionado aún más por Mutch et al. [14], introduciendo escasez y lateral inhibición. Aquí, la agrupación máxima se realiza en parches superpuestos con una superposición factor de dos en una vecindad espacial y a través de escalas.

Aunque todos estos métodos se parecen al concepto de simple y complejo celdas, se diferencian por el tipo de entradas, el método de extracción de características, algoritmo de entrenamiento y el clasificador utilizado. Además, cada método emula células complejas ligeramente diferentes. Debido a esas variaciones, un análisis exhaustivo de cómo Es imposible que una elección particular de este componente afecte el rendimiento general. En para comparar empíricamente la influencia única de diferentes funciones de agregación, Elegimos una arquitectura de modelo fija, que se describe en la siguiente sección.

3 Arquitectura modelo

Esta sección describe la arquitectura de extracción y clasificación de características.

sistema, así como el procedimiento de entrenamiento utilizado en nuestros experimentos. Nosotros elegimos realizar nuestras evaluaciones en el marco de un sistema neuronal convolucional Red (CNN). Las CNN han logrado resultados de última generación para el reconocimiento de dígitos escritos a mano [23] y para la detección de rostros [17]. Están desplegados en sistemas comerciales para leer cheques [10] y ofuscar rostros y matrículas en Google StreetView [5].

3.1 Modelo básico

Las CNN son representantes de la arquitectura de múltiples etapas de Hubel-Wiesel, que extraer características locales en alta resolución y combinarlas sucesivamente en características más complejas a resoluciones más bajas. La pérdida de información espacial es compensado por un número cada vez mayor de mapas de características en las capas superiores. CNN constan de dos tipos diferentes de capas: capas convolucionales (capas C), que

se asemejan a las células simples y las capas de agrupación (capas P), que modelan el comportamiento de las células complejas. Cada capa convolucional realiza una operación de convolución 2D discreta en su imagen fuente con un núcleo de filtro y aplica una función de transferencia no lineal. Las capas de agrupación reducen el tamaño de la entrada al resumir las neuronas de una pequeña vecindad espacial. Nuestra elección de una CNN está motivada en gran medida por el hecho de que la operación realizada por las capas de agrupación es fácilmente intercambiable sin modificaciones en la arquitectura. La arquitectura general se muestra en la Figura 1 y es idéntica a los modelos de LeCun et al. [11] y Ahmed et al. [1].

3.2 Capas convolucionales Los cálculos

para el paso directo y la retropropagación en la capa convolucional siguen el procedimiento estándar en la literatura y tienen filtros entrenables y un sesgo entrenable por mapa de características. Se aplica una función tangente hiperbólica a las activaciones en esta capa. Nuestros experimentos han demostrado que una conexión escasa entre mapas de características no mejora el rendimiento del reconocimiento en comparación con mapas de características completamente conectados siempre que el número de parámetros sea igual. Por lo tanto, en una capa convolucional, cada mapa está conectado a todas sus características anteriores.

3.3 Capas de agrupación

El propósito de las capas de agrupación es lograr invariancia espacial reduciendo la resolución de los mapas de características. Cada mapa de características agrupado corresponde a un mapa de características de la capa anterior. Sus unidades combinan la entrada de un pequeño parche de unidades de $n \times n$, como se indica en la Figura 1. Esta ventana de agrupación puede ser de tamaño arbitrario y las ventanas pueden superponerse.

Evaluamos dos operaciones de agrupación diferentes: agrupación máxima y submuestreo. La función de submuestreo

$$a_j = \tanh(\beta \sum_{i=1}^{n \times n} u_{ij} + b) \quad (1)$$

$N \times N$

toma el promedio de las entradas, lo multiplica con un escalar entrenable β , agrega un sesgo entrenable b y pasa el resultado a través de la no linealidad. La función de agrupación máxima

$$a_j = \max_{i=1}^{n \times n} (u_i(n, n)) \quad (2)$$

$N \times N$

aplica una función de ventana $u(x, y)$ al parche de entrada y calcula el máximo en la vecindad. En ambos casos, el resultado es un mapa de características de menor resolución.

3.4 Retropropagación Todas

las capas del modelo CNN se entrenan utilizando el algoritmo de retropropagación. Para la propagación de errores y la adaptación del peso en capas totalmente conectadas, convolucionales y de submuestreo, seguimos el procedimiento estándar. En las capas de agrupación máxima, las señales de error solo se propagan a la posición en $\arg \max_{i=1}^{n \times n} (u_i(n, n))$. Por lo tanto, los mapas de error en las capas de agrupación máxima son escasos. Si se utilizan ventanas de agrupación superpuestas, podría ser necesario acumular varias señales de error en una unidad.



Fig. 2. Algunos ejemplos de imágenes preprocesadas del conjunto de datos Caltech-101.



Fig. 3. Imágenes del conjunto de datos uniforme normalizado NORB.

4 resultados experimentales

El propósito de nuestros experimentos es triple: (1) mostrar que la agrupación máxima es superior al submuestreo, (2) determinar si las ventanas de agrupación superpuestas pueden mejorar el rendimiento del reconocimiento y (3) encontrar funciones de ventana adecuadas. Todos los experimentos se realizaron con la misma inicialización aleatoria a menos que se indique lo contrario.

Para acelerar la capacitación, implementamos la arquitectura CNN en unidades de procesamiento de gráficos (GPU) utilizando el marco de programación CUDA de NVIDIA [16]. Las operaciones de convolución utilizan rutinas amablemente proporcionadas por Alex Krizhevsky¹. La mayoría de las demás operaciones se aceleran con nuestra biblioteca CUV disponible públicamente [13]. Para el aprendizaje por mini lotes, con solo unos pocos patrones procesados en paralelo, logramos una aceleración de aproximadamente dos órdenes de magnitud en comparación con nuestra implementación de CPU.

4.1 Conjuntos de datos

Evaluamos diferentes operaciones de agrupación en los conjuntos de datos Caltech-101 [4] y NORB [11]. Varios autores han publicado tasas de reconocimiento con otras arquitecturas CNN para ambos conjuntos de datos.

El conjunto de datos Caltech-101 consta de 101 categorías de objetos y una categoría de fondo. Hay un total de 9144 imágenes de diferentes tamaños de aproximadamente 300×300 píxeles. Preprocesamos las imágenes ajustándolas en un marco de imagen de 140×140 , conservando su relación de aspecto. El relleno se rellenó con la media de la imagen para cada canal de color. Difuminamos el borde de la imagen en el relleno para eliminar los efectos secundarios causados por el borde de la imagen, como se muestra en la Figura 2. Las imágenes resultantes se normalizan por canal para tener una media de cero y una varianza de uno para acelerar el aprendizaje.

Seguimos el protocolo de experimento común en la literatura para Caltech-101, que consiste en elegir aleatoriamente 30 imágenes de cada categoría para entrenamiento y uso.

¹ <http://www.cs.utoronto.ca/~kriz>

el resto para pruebas. La tasa de reconocimiento se mide para cada clase y reportamos el promedio de las 102 clases.

Además, realizamos experimentos en el conjunto de datos uniforme normalizado NORB, que consta de solo cinco categorías de objetos. El conjunto de entrenamiento y prueba contiene cada uno cinco instancias de objetos de juguete para cada categoría, fotografiados desde diferentes ángulos y bajo diferentes iluminaciones. Cada patrón consta de un par binocular de imágenes en escala de grises de 96×96 , con un total de 24.300 patrones de entrenamiento y la misma cantidad de patrones de prueba.

4.2 Agrupación máxima versus submuestreo

Para mantener nuestros resultados comparables, diseñamos deliberadamente las redes para que se parecieran a la de Huang y LeCun [7] para NORB y a la de Ahmed et al. [1] para Caltech-101.

Para NORB, la capa de entrada con dos mapas de características de tamaño 96×96 va seguida de una capa convolucional C1 con filtros de 5×5 y 16 mapas de tamaño 92×92 . P2 es una capa de agrupación de 4×4 , que reduce el tamaño a 23×23 . La capa convolucional C3 emplea filtros de 6×6 y tiene 32 mapas con dimensiones de 18×18 píxeles. La capa de agrupación P4 con ventanas de agrupación de 3×3 produce mapas de características de 6×6 que están completamente conectados a 100 neuronas ocultas. La capa de salida recibe información de las 100 neuronas y codifica la clase de objeto con 5 neuronas. Esta red de seis capas se muestra en la Figura 1.

Se entrenaron dos variaciones de la red: en el primer caso, las capas de agrupación realizaron una operación de submuestreo, en el segundo caso, esto se reemplazó con una agrupación máxima. Tenga en cuenta que la red de submuestreo tiene un poco más de parámetros debido al escalar y al sesgo entrenable, como se describe en la Sección 3. Después de 1000 épocas de entrenamiento de retropropagación con minilotes de 60 patrones, existe una discrepancia sorprendente entre las tasas de reconocimiento. Para cada red, se realizaron cinco pruebas con diferentes inicializaciones de peso y reportamos la tasa de error promedio y la desviación estándar. La red de submuestreo logra una tasa de error de prueba del 7,32 % ($\pm 1,27$ %), en comparación con el 5,22 % ($\pm 0,52$ %) para la red de agrupación máxima.

Para Caltech-101, la capa de entrada consta de tres mapas de características de tamaño 140×140 , seguidas de una capa convolucional C1 con filtros de 16×16 y 16 mapas de características. La siguiente capa de agrupación P2 reduce los mapas de 125×125 con ventanas de agrupación no superpuestas de 5×5 a un tamaño de 25×25 píxeles. La capa convolucional C3 con filtros 6×6 consta de 128 mapas de características de tamaño 20×20 . La capa de agrupación P4 utiliza una ventana de agrupación no superpuesta de 5×5 . Las neuronas de esos 128 mapas de características de tamaño 4×4 están completamente conectadas a las 102 unidades de salida.

Después de 300 épocas utilizando el algoritmo de aprendizaje por lotes RPROP [19], evaluamos el rendimiento de reconocimiento normalizado para Caltech-101. Al utilizar una operación de submuestreo, la red logró una tasa de error de prueba del 65,9%. Sustituir esto con una operación de agrupación máxima arrojó una tasa de error del 55,6%.

Tanto para NORB como para Caltech-101, nuestros resultados indican que las arquitecturas con una operación de agrupación máxima convergen considerablemente más rápido que aquellas que emplean

	NORB				Caltech-101			
	set de tren		set de prueba		set de tren		set de prueba	
superposición	0,00%	6,40%	1,28%	52,29%	sin			
2 píxeles superpuestos	0,00%	6,48%	2,29%	52,74%	4			
píxeles superpuestos	0,00%	6,37%	3,92%	52,42%	6 píxeles			
superpuestos	0,01%	7,27%	4,55%	53,82%	8 píxeles			
superpuestos	0 0,00%	6,84%	7,43%	55,79%	10 píxeles se			
superponen	0,01%	7,21%	10,17%	57,32%				

Tabla 1. Tasas de reconocimiento en NORB normalizado uniforme (después de 300 épocas) y Caltech-101 (después de 400 épocas) para redes con diferentes cantidades de superposición en las capas de agrupación máxima.

una operación de submuestreo. Además, parecen ser superiores en la selección de características invariantes y mejoran la generalización.

4.3 Ventanas de agrupación superpuestas

Para evaluar cómo el tamaño del paso de las ventanas de agrupación superpuestas afecta las tasas de reconocimiento, básicamente utilizamos las mismas arquitecturas que en la sección anterior.

Sin embargo, ajustar el tamaño del paso cambia el tamaño de los mapas de características y con ello el número total de parámetros entrenables, así como la relación entre pesos completamente conectados y pesos compartidos. Por lo tanto, estamos aumentando el tamaño de los mapas de características de entrada en consecuencia, colocando el patrón de entrada en el centro de un mapa de características y rellenándolo con ceros. En la arquitectura NORB, por ejemplo, los mapas de características de entrada tienen un tamaño de 106×106 , si se elige un tamaño de paso de dos.

La Tabla 1 enumera las tasas de reconocimiento para diferentes tamaños de paso en ambos conjuntos de datos. El rendimiento se deteriora si se aumenta el tamaño del paso. Esto podría deberse al hecho de que no se gana información si las ventanas de agrupación se superponen. Los máximos en las regiones de ventanas superpuestas simplemente se duplican en la siguiente capa y los píxeles vecinos están más correlacionados.

4.4 Funciones de ventana

Pequeñas variaciones de la imagen de entrada y desplazamientos más allá del borde de la ventana de agrupación pueden cambiar drásticamente la representación. Por esta razón, experimentamos con ventanas de agrupación superpuestas y más suaves. Las funciones de ventana se utilizan a menudo para suavizar una señal de entrada en aplicaciones de procesamiento de señales. Hemos evaluado cuatro funciones de ventana diferentes, como se muestra en la Tabla 2.

Nuevamente, la arquitectura de red para esos experimentos fue similar a la de las secciones anteriores. Para el conjunto de datos NORB, la red se modificó para recibir entradas de 128×128 y grupos P2 desde una ventana de 12×12 con una superposición de 8 píxeles. Las unidades en P4 reciben información de ventanas de 9×9 , que se superponen en 6 píxeles. Por lo tanto, si se elige una función de ventana rectangular pequeña, esto equivale a la red no superpuesta. De manera similar, para Caltech-101, la entrada se rellenó a







	Sin superposición	Cono rectangular	Pirámide	Triángulo	Binomio	
						
Error de prueba NORB 5,56 %	Error de	5,83%	6,29%	6,28%	10,92%	12,15%
prueba Caltech-101 52,25 %		51,95%	69,95%	73,16%		

Tabla 2. Tasas de error de prueba para NORB (después de 500 épocas de entrenamiento) y para Caltech-101 (después de 600 épocas). Aplicar una función de ventana a un vecindario superpuesto es consistentemente peor que usar ventanas agrupadas que no se superpongan.

230 × 230 y las capas P2 y P4 se agrupan desde ventanas de 15 × 15 y 18 × 18, respectivamente.

Como se muestra en la Tabla 2, ninguna de las funciones de ventana mejoró significativamente las tasas de reconocimiento. La función de ventana binomial y la función de ventana triangular incluso producen resultados sustancialmente peores que una función de ventana rectangular.

4.5 Resultados de otros conjuntos de datos

Hasta donde sabemos, no se han publicado resultados que utilicen operaciones de agrupación máxima en el conjunto de datos MNIST de dígitos escritos a mano [10] y el conjunto de datos desordenados NORB [11]. Por lo tanto, aplicamos nuestro modelo de red con ventanas de agrupación máximas no superpuestas a ambos conjuntos de datos.

Para MNIST, logramos un error de prueba del 0,99 % con una arquitectura bastante superficial. Aquí, la capa de entrada de 28 × 28 fue seguida por una capa convolucional C1 con filtros de 9 × 9 y 112 mapas de características. En la siguiente capa de agrupación máxima P2, cada unidad recibe información de una ventana de 5 × 5 y está conectada a cada una de las diez neuronas de salida. Entrenamos esta arquitectura para 60 épocas de actualizaciones en línea con retropropagación.

El conjunto de datos desordenados de NORB se evaluó con la siguiente arquitectura : los patrones de entrada estéreo de tamaño 108 × 108 están convolucionados en C1 con filtros de 5 × 5 en 16 mapas de características y se agrupan como máximo en P2 con ventanas de 9 × 9 y una superposición. de 4 píxeles. Están convolucionados con filtros de 8 × 8 en C3 para producir 32 mapas de características de tamaño 13 × 13. La capa P4 reduce el tamaño a 5 × 5 píxeles agrupando ventanas de 5 × 5 con una superposición de 3 píxeles. Dos capas completamente conectadas con 100 neuronas y seis neuronas de salida concluyen la arquitectura. Entrenar este modelo durante siete épocas con aprendizaje en línea y minilotes de 60 patrones arrojó una tasa de error de prueba del 5,6%. Hasta donde sabemos, este es el mejor resultado publicado sobre este conjunto de datos hasta ahora.

También mejoramos nuestros resultados en NORB uniforme normalizado con algunas pasadas de refinamiento utilizando un esquema de énfasis. Durante estos pases se entrenaron con mayor frecuencia patrones difíciles con errores superiores a la media. En este caso, logramos una tasa de error de prueba del 4,57 %, que supera el 5,2 % informado por Nair y Hinton [15], a pesar de que utilizaron datos adicionales sin etiquetar.

5. Conclusión

Hemos demostrado que una operación de agrupación máxima es muy superior para capturar invariancias en datos similares a imágenes, en comparación con una operación de submuestreo. Para varios conjuntos de datos, los resultados del reconocimiento con una arquitectura equivalente mejoran enormemente con respecto a las operaciones de submuestreo. En NORB uniforme normalizado (4,57 %) y NORB jittered-cluttered (5,6 %) incluso logramos los mejores resultados publicados hasta la fecha.

Sin embargo, el uso de ventanas de agrupación superpuestas y más fluidas no mejora las tasas de reconocimiento. En trabajos futuros, investigaremos más a fondo para encontrar funciones de ventana adecuadas. Los histogramas (como en [3, 12]) pueden verse como un tercer tipo de operación de agrupación que aún no se ha evaluado en profundidad. La combinación de dichas operaciones de histograma con redes neuronales convolucionales podría mejorar aún más las tasas de reconocimiento en tareas de visión.

Referencias

1. A. Ahmed, K. Yu, W. Xu, Y. Gong y E. Xing. Entrenamiento de modelos jerárquicos de reconocimiento visual feed-forward mediante aprendizaje por transferencia a partir de pseudotareas. *Visión por computadora – ECCV 2008*, páginas 69–82.
2. Sven Behnke. Redes neuronales jerárquicas para la interpretación de imágenes, volumen 2766 de *Lecture Notes in Computer Science*. Springer-Verlag Nueva York, Inc., junio de 2003.
3. Navneet Dalal y Bill Triggs. Histogramas de gradientes orientados para humanos Detección. En *CVPR*, páginas 886–893, 2005.
4. L. Fei-Fei, R. Fergus y P. Perona. Aprendizaje de modelos visuales generativos a partir de algunos ejemplos de entrenamiento: un enfoque bayesiano incremental probado en 101 categorías de objetos. *Visión por computadora y comprensión de imágenes*, 106(1):59–70, 2007.
5. A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven y L. Vincent. Protección de privacidad a gran escala en Google Street View. California, UEA, 2009.
6. Kunihiro Fukushima. Un modelo de red neuronal para la atención selectiva en visual. reconocimiento de patrones. *Cibernética biológica*, 55(1):5–15, octubre de 1986.
7. Fu-Jie Huang y Yann LeCun. Aprendizaje a gran escala con svm y redes convolucionales para categorización genérica de objetos. En *Proc. Jornada de Visión por Computador y Reconocimiento de Patrones (CVPR'06)*. Prensa IEEE, 2006.
8. DH Hubel y TN Wiesel. Campos receptivos de neuronas individuales en la corteza estriada del gato. *La Revista de Fisiología*, 148(3):574, 1959.
9. Svetlana Lazebnik, Cordelia Schmid y Jean Ponce. Más allá de muchas funciones: coincidencia de pirámides espaciales para reconocer categorías de escenas naturales. En *CVPR (2)*, páginas 2169–2178. Sociedad de Computación IEEE, 2006.
10. Y. LeCun, L. Bottou, G. Orr y K. Müller. BackProp eficiente. En *Redes neuronales: trucos del oficio*. Springer, 1998.
11. Y. LeCun, F. Huang y L. Bottou. Métodos de aprendizaje para el reconocimiento de objetos genéricos con invariancia de pose e iluminación. En *Actas de CVPR'04*. Prensa IEEE, 2004.
12. David G. Lowe. Características de imagen distintivas a partir de puntos clave invariantes de escala. *Revista Internacional de Visión por Computadora*, 60:91–110, 2004.

13. A. Muller, H. Schulz y S. Behnke. Características topológicas en RBM conectados localmente. En Proc. Conferencia conjunta internacional sobre redes neuronales (IJCNN 2010), 2010.
14. J. Mutch y DG Lowe. Reconocimiento de objetos multiclase con funciones localizadas y dispersas . En la Conferencia de la IEEE Computer Society sobre visión por computadora y reconocimiento de patrones, volumen 1, páginas 11 a 18, junio de 2006.
15. V. Nair y G. Hinton. Reconocimiento de objetos tridimensionales con redes de creencias profundas. Avances en Sistemas de procesamiento de información neuronal, 2010.
16. Corporación Nvidia. Guía de programación CUDA 3.0, febrero de 2010.
17. M. Osadchy, Y. LeCun y M. Miller. Detección sinérgica de rostros y estimación de poses con modelos basados en energía. Revista de investigación sobre aprendizaje automático, 8:1197–1215, mayo de 2007.
18. Marc'Aurelio Ranzato, Fu-Jie Huang, Y-Lan Boureau y Yann LeCun. Aprendizaje no supervisado de jerarquías de características invariantes con aplicaciones al reconocimiento de objetos. En Proc. Conferencia sobre visión por computadora y reconocimiento de patrones (CVPR'07). Prensa IEEE , 2007.
19. Martin Riedmiller y Heinrich Braun. RPROP – Descripción y detalles de implementación. Informe técnico, Universidad de Karlsruhe, enero de 1994.
20. M. Riesenhuber y T. Poggio. Modelos jerárquicos de reconocimiento de objetos en la corteza. Nature Neuroscience, 2:1019–1025, 1999.
21. T. Serre, L. Wolf y T. Poggio. Reconocimiento de objetos con características inspiradas en la corteza visual. En Conferencia de la IEEE Computer Society sobre visión por computadora y reconocimiento de patrones, 2005. CVPR 2005, volumen 2, 2005.
22. C. Siagian y L. Itti. Clasificación rápida de escenas de inspiración biológica utilizando características compartidas con la atención visual. Transacciones IEEE sobre análisis de patrones e inteligencia artificial, 29(2):300, 2007.
23. Patrice Y. Simard, Dave Steinkraus y John C. Platt. Mejores prácticas para redes neuronales convolucionales aplicadas al análisis de documentos visuales. En Conferencia internacional sobre análisis y reconocimiento de documentos (ICDAR), IEEE Computer Society, Los Alamitos, páginas 958–962, 2003.