High-Dimensional Learning and Deep Neural Networks

Sebastián Oviedo, Fabián Rincón

July 3, 2022

1 Introducción

La arquitectura que se implementa en las redes neuronales convolucionales (CNN) es particularmente reconocida por tener buenos resultados para datos con alta dimensionalidad [BGKP21]. Como consecuencia se han usado para intentar resolver problemas de clasificación de imagen y procesamiento de audio, entre otros. Se trata de una arquitectura multicapa en la que se efectúan una serie de operadores lineales junto con unos operadores no lineales [Mal16]. Esto causa una complejidad inherente que solo permite entender en cierta medida las capas más superficiales de la red y desconociendo el comportamiento y aporte de las capas más ocultas, lo que dificulta la explicación e interpretación de cada capa lo que es algunas ocaciones es de vital importancia para los usuarios finales. Por esta razón Stephane Mallat propone una aproximación conceptual para clarificar y entender las redes neuronales, fundamentada en conceptos matemáticos relativamente conocidos y comprensibles.

2 Entendiendo las CNN

Los algoritmos de aprendizaje buscan una función $\hat{f}(x)$ que se aproxime a la funciòn objetivo $f(x)$ a partir de una muesra de q ejemplos $\{x^i, f(x^i)\}_{i\leqslant q}$ para $x = (x(1), ..., x(d))$. El valor de d es alto, representando la alta dimensionalidad. En efecto la alta dimensionalidad es la que pone en discusión la complejidad de las CNN permitiendo obtener una línea conceptual que parte de allí.

2.1 Los tres principios fundamentales de los algoritmos de aprendizaje de máquina

Como vimos al inicio del curso, en cualquier ejercicio de aprendizaje de máquina, es necesario tener en cuenta 3 conceptos importantes. 1) el objetivo del problema; es decir, identificar si estamos en un problema de clasificación o regresión, 2) definir la función de pérdida a tener en cuenta en todo el proceso de aprendizaje y, 3) el algoritmo de optimización usado para encontrar los párametros que minimizan la funcion de pérdida.

En la Figura 1 podemos observar una representación gráfica de los componentes de una CNN, donde se identifican los datos de entrada en la primera capa, tambien encontramos algunas de las operaciones

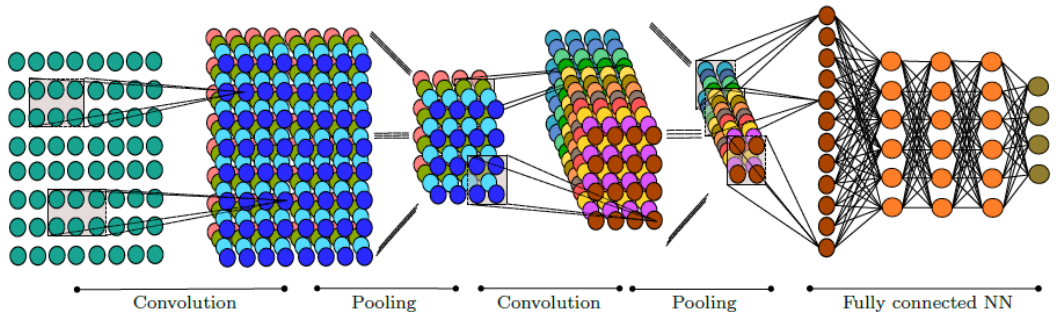

Figure 1: Ilustración de una red neuronal convolucional [BGKP21]

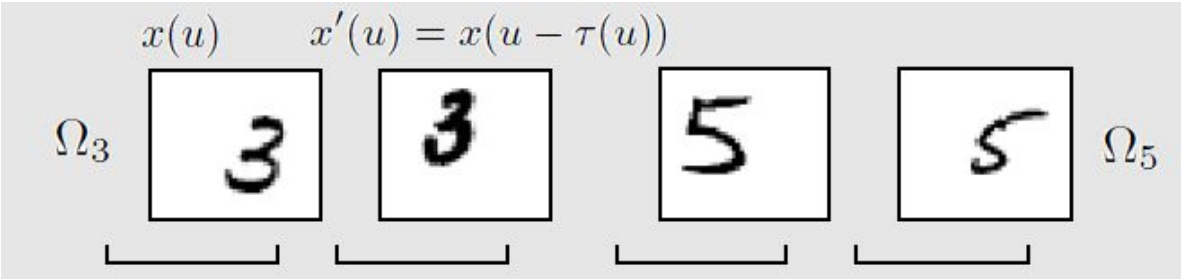

Figure 2: Ejemplo de acción para dos dígitos [BGKP21]

más utilizadas como las convoluciones y funciones de activación, las cuales se repiten varias veces durante cada una de las capas de las estructuras

2.2 Reducción de dimensión

Es importante resaltar que desde este enfoque es necesario tener en consideración algunas estrategias para reducir la dimensionalidad. Una es la *Separación*, que permite distinguir dos elementos apropiadamente después de proyectarlos en espacios de dimensión menor y la segunda, la *linealización* que se presenta como una estrategia para reducir la dimensión a partir de la construcción de espacios de baja de dimensión. En todo caso es esencial que la *variabilidad* perdure para que no exista la pérdida de información.

- Separación: En este caso se quiere reducir la dimensión de x a un vector de $\Phi(x)$.Consideremos que la separabilidad es la propiedad en la que si $f(x) \neq f(x')$ implica que $\Phi(x) \neq \Phi(x')$ ç. Cuando esto sucede decimos que Φ separa a f. Ahora bien, esta funciòn que permita realizar la proyección debe tener una porpiedad adicional con la cual se garantiza que exista tal separaciòn pero que garantize o bien la de la Entonces es necesario encontrar una función $\Phi(x)$ la cual permite conservar en cierto sentido la continuidad, importante para problemas de regresión y para la identificaciòn de clases. Esta propoedad adicional se denomina Lipschitz que, en terminos generales se describe de la siguiente manera:

$$\| \Phi(x) - \Phi(x') \| \geq \epsilon |f(x) \neq f(x')| \tag{1}$$

- Linealización: Consiste en hacer un cambio de variable apropiado $\Phi(x) = \{\phi_k(x)\}_{k \leqslant d}$. En este caso decimos que Φ separa linealmente a f si $f(x)$ está bien aproximada por una proyección, es decir

$$\hat{f}(x) = \langle \Phi(x), w \rangle = \sum w_k \phi_k(x) \tag{2}$$

2.3 Propiedades localmente invariables

La necesidad de encontrar la función $f(x)$ a partir de un conjunto de datos hace que tengamos la necesidad de encontrar ciertas propiedades que permiten caracterizar la variabilidad lo suficientemente bien para distinguir apropiadamente distintas clases. Una es la traslación y la otra es una deformación continua (difeimorfismo) figura 2.

Estas a su vez se tienen la propiedad en este caso de ser localmente invariantes, esto es si g es la acción sobre x y si definimos una constante C_x como una medida de qué es lo local con ($|g|_G \leq C_x$) se tiene que $f(g.x) = f(x)$, en otras palabras la acción no actúa dentro de la influencia definidad por C_x. La acción g puede ser la traslación o el difeimorfismo nombrados anteriormente.

Estas propiedades deben permanecer despues de aplicar $\Phi(x)$ por lo que también es necesario que se cumpla la siguiente condición para $\Phi(x)$, denominada Lipschitz continua.

$$\exists C > 0, \|\Phi(g.x) - \Phi(x)\| < C|g|_G \|x\| \tag{3}$$

Classification Errors

Training size	Conv. Net.	Scattering
50000	0.4%	0.4%

LeCun et. al.

Figure 3: Comparación de resultados de las dos metodologías

3 Comparación de las CNN y las Waveles de Mallat

Como se mencionó anteriormente, las redes convolucionales han sido ampliamente usadas en tarea de procesamiento de imágenes y audio, tambien se mencionó que que se debe tener en cuenta los 3 principios fundamentales de las CNN; sin embargo, el uso de este enfoque dificulta la interpretación en cada capa de la red dadas las no linealidades usandas en las funciones de activación. Por otro lado, la propuesta de Mallat nos ofrece una oportunidad para abordar los mismos problemas de las CNN usando las transformaciones definidas anteriormente, para lo cual, no tenemos necesidad de considerar la función de pérdida, el algoritmo de optimización y todos los problemas de aprendizaje como los requerimientos de computo.

3.1 Aplicación

En la Figura 3 se presentan los datos ampliamente conocidos para clasificación de dígitos MNIST, tambien se presentan los resultados de las 2 metodologías mencionadas anteriormente: 1) la aplicación de las CNN y 2) el uso de las Wavelets de Mallat. Como se puede observar las 2 metodologías obtienen los mismos resultados en términos de error de clasificación, sin embargo, como se mencionó previamente, el uso de las Wavelets no requirió los mismos insumos de las CNN y, más importante aún, perimite obtener una interpretación directa de los resultados.

A manera de conclusión, es importante mencionar que, así como hoy en día se vive un auge en el uso de las CNN para solucionar los problemas relacionados con imágenes por sus resultados satisfactorios. En el futuro cercano se verá una tendencia hacia métodos más sencillos e interpretables como las Wavelets.

References

[BGKP21] Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning, 2021.

[Mal16] Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, apr 2016.