Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Correspondencia entre respuesta electrofisiológica ante estímulos visuales y activaciones en redes neuronales convolucionales

Tesis de Licenciatura en Ciencias de Datos

Agustin de Otazua

Director: Enzo Tagliazucchi
Codirector: Eric Lützow Holm
Buenos Aires, 2023

A mi familia, que siempre me acompañó en todo.

ABSTRACT

El sistema visual de los primates procesa los estímulos visuales en forma jerárquica. En su etapa inicial, la corteza visual primaria responde a características de bajo nivel, como lo son los segmentos orientados. Progresivamente, las siguientes regiones de la corteza procesan patrones cada vez más complejos, tales como el color, el movimiento y las formas compuestas, alcanzando, por último, a la representación de objetos o conceptos específicos en la escena visual. Las redes neuronales convolucionales, inspiradas en el sistema visual humano, también presentan un procesamiento jerárquico de la información visual. Esto ha llevado a sugerir que las redes neuronales convolucionales pueden usarse como modelo para comprender el proceso de información en el sistema visual. En esta tesis, investigamos la correspondencia temporal entre ambos sistemas, partiendo de la hipótesis de que las respuestas más tempranas de electroencefalografía estarán relacionadas con las capas más tempranas de una red neuronal convolucional que procesa el mismo estímulo, y viceversa para las componentes más tardías. Para ello, analizamos un conjunto de registros electrofisiológicos en individuos percibiendo un gran conjunto de imágenes de forma secuencial. Por otra parte, obtenemos las activaciones por capa de una red neuronal convolucional preentrenada usando las mismas imágenes como entrada. Exploramos dos técnicas distintas para analizar la correspondencia: primero mediante un análisis de similitud representacional entre ambos sistemas, y luego mediante la predicción de las activaciones cerebrales en base a las activaciones en la red neuronal. Verificamos que en algunos casos las primeras capas de la red se corresponden mayormente con etapas tempranas de las señales electrofisiológicas, mientras que las capas avanzadas se corresponden con etapas más tardías, apoyando la hipótesis propuesta.

Palabras claves: Neurociencia, Visión, EEG, CNN, Machine-learning, Deep-learning.

Índice general

1.. Introducción 1
1.1. Visión en humanos y otros primates 1
1.2. Redes neuronales convolucionales 2
1.2.1. Capas convolucionales 3
1.2.2. Capas de pooling 4
1.2.3. Capas densas 4

2.. Activaciones biológicas y artificiales 6
2.1. Dataset de imágenes 6
2.2. Activaciones electrofisiológicas 6
2.3. Activaciones de la CNN 7

3.. Análisis de similitud representacional - matrices de disimilitud 9
3.1. Descripción 9
3.2. Motivación 9
3.3. RDM - EEG 9
3.4. RDM - CNN 10

4.. Correspondencia entre respuesta de EEG y activaciones en CNN 12
4.1. Metodología 13
4.2. Resultados 13
4.2.1. Con todos los individuos 13
4.2.2. Resultados con individuos seleccionados 14

5.. Predicción de respuestas EEG a partir de activación en CNN 18
5.1. Metodología 18
5.2. Resultados 19

6.. Conclusiones 21
6.1. Resultados obtenidos 21
6.2. Comparación con otros estudios 21
6.3. Limitaciones de EEG 22
6.4. Limitaciones de las CNN 22
6.5. Trabajo a futuro 22

1. INTRODUCCIÓN

En este capítulo presentamos una introducción al procesamiento de información visual en la corteza cerebral, y a la arquitectura de redes neuronales convolucionales aplicadas a la visión artificial o por computadora. En particular, indicaremos algunos paralelismos entre ambos sistemas, lo cual sugiere que las redes neuronales convolucionales podrían utilizarse como un modelo computacional del procesamiento de la información visual en el cerebro.

1.1. Visión en humanos y otros primates

La corteza visual primaria (V1) es una región del cerebro ubicada en el lóbulo occipital, la cual juega un papel fundamental en el procesamiento de la información visual. Esta región recibe señales visuales de la retina a través del nervio óptico, pasando por el tálamo, y las procesa para extraer información sobre la escena visual. Dicha información incluye la detección de colores, movimientos, formas, y objetos específicos.

El experimento de Hubel y Wiesel [1] fue un estudio revolucionario realizado en la década de 1950 que ayudó a comprender el procesamiento de información en V1. Utilizando electrodos implantados en gatos y monos, los investigadores registraron la actividad de las células en la corteza visual mientras los animales eran expuestos a diferentes estímulos visuales (Fig. 1.1, izquierda). Descubrieron que las células en V1 respondían selectivamente a diferentes características visuales, como la orientación, el movimiento y la forma de los estímulos. En partícular, mostraron que las neuronas de V1 responden a segmentos con orientaciones específicas. Por ejemplo, algunas células mostraban una mayor actividad cuando se presentaban barras horizontales, mientras que otras respondían mejor a barras verticales. En resumen, encontraron que todas las neuronas maximizaban su respuesta ante segmentos con una orientación determinada.

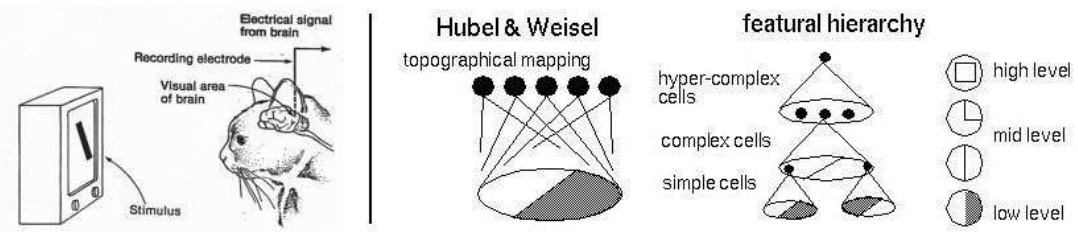


Fig. 1.1: Diagrama del experimento de Hubel y Wiesel. A la izquierda, se ilustra el procedimiento llevado a cabo para presentar los estímulos y registrar las señales de la corteza visual [1]. A la derecha, se muestra el modelo de Hubel y Wiesel, según el cual la información se procesa de forma jerárquica, con las respuestas de neuronas que responden a información de bajo nivel siendo integradas en neuronas que responden a patrones cada vez más complejos en las imágenes [2].

Estos hallazgos muestran, por lo tanto, que V1 está especializada en el procesamiento de un aspecto específico de la información visual, dado que células individuales tienen preferencias distintas en cuanto a las características visuales que detectan. Después de V1, la información visual se procesa en varias áreas adicionales de la corteza visual, conocidas

como áreas visuales secundarias y asociativas. Estas áreas se encuentran en diferentes regiones del cerebro, y en ellas se encuentran células que responden ante patrones más complejos; por ejemplo, las áreas visuales secundarias, como el área visual secundaria (V2) y el área visual secundaria dorsal (V3), reciben información de V1 y desempeñan un papel importante en la integración de características visuales más complejas, como el color, la forma y el movimiento.

En base a sus experimentos, Hubel y Wiesel propusieron un modelo para explicar de forma integral el funcionamiento de la visión en la corteza cerebral [2]. En este modelo, las neuronas de V1 extraen información sobre la orientación de segmentos, y proyectan su respuesta a neuronas que integran esta información de múltiples neuronas de V1, y que por lo tanto pueden responder a combinaciones más complejas de segmentos orientados. A su vez, estas neuronas envían la información para ser integradas posteriormente en otras neuronas, resultando en un proceso iterativo jerárquico donde las neuronas cada vez son capaces de generar respuestas más y más selectivas ante ciertos patrones en la escena visual (Fig. 1.1, derecha).

Este proceso culmina cuando la información visual alcanza neuronas que se encuentran fuera de las áreas visuales secundarias, en regiones de la corteza denominadas áreas visuales asociativas. Estas áreas, tales como el área visual temporal (V4) y el área visual parietal (V5 o MT), están involucradas en la identificación de objetos, la percepción de la profundidad y el seguimiento de objetos en movimiento. En particular, en la corteza temporal inferior (IIT) se encuentran neuronas que disparan selectivamente ante la presencia de objetos o conceptos específicos en la escena visual [3].

Finalmente, la información visual se proyecta a otras regiones del cerebro involucradas en otras funciones cognitivas, tales como la memoria, la atención y la toma de decisiones.

1.2. Redes neuronales convolucionales

Las *redes neuronales convolucionales* (CNN, por sus siglas en inglés) son un tipo de arquitectura de redes neuronales artificiales que se utilizan principalmente en tareas de visión por computadora, como reconocimiento de imágenes y segmentación de objetos. Fueron desarrolladas en la década de 1980, pero su popularidad y avances significativos se produjeron en los últimos años.

La inspiración para las CNNs provino, de forma indirecta, de la investigación en neurociencia y la aplicación de esas ideas en el campo de la inteligencia artificial. Un precursor de las CNN es el modelo *Neocognitron*, desarrollado por Kunihiko Fukushima, el cual intenta capturar el procesamiento jerárquico de la información en la corteza cerebral [4]. Las CNN se basaron a su vez en este modelo, adoptando la idea de que las células individuales responden a regiones específicas de la imagen en vez de toda la imagen a la vez, tal como sería el caso de la redes densas, resultando en una arquitectura que posee menos parámetros para ajustar y es invariante ante traslaciones [5].

Uno de los hitos clave en el desarrollo de las CNNs fue la propuesta del algoritmo de *backpropagation* en 1986 [6], que permitió el entrenamiento eficiente de redes neuronales con múltiples capas, allanando el camino para el desarrollo de arquitecturas más profundas y complejas, como las CNNs. En 1989, Le Cun et al. [7] utilizaron dicho algoritmo para entrenar una de las primeras CNNs en el reconocimiento de caracteres escritos a mano.

En términos de aplicaciones usuales, estas redes han demostrado ser muy efectivas en una variedad de tareas de visión por computadora, como:

- Reconocimiento de objetos: las CNNs pueden identificar y clasificar objetos en imágenes, como perros, gatos, automóviles, etc.
- Detección de rostros: se utilizan en sistemas de reconocimiento facial para detectar y reconocer rostros en imágenes y videos.
- Reconocimiento de caracteres: se utilizan en aplicaciones de reconocimiento óptico de caracteres (OCR) para leer texto impreso o manuscrito [7].
- Conducción autónoma: las CNNs se utilizan en sistemas de visión por computadora para identificar y seguir objetos en tiempo real, como peatones, vehículos y señales de tráfico.

En cuanto a los elementos de una CNN, existen tres tipos principales de capas: *convolucional*, de *pooling* y *densa* [8]. La cantidad, tamaño y disposición de cada una determinan la *arquitectura* de la red. Un ejemplo de esta arquitectura se muestra en la figura 1.2.

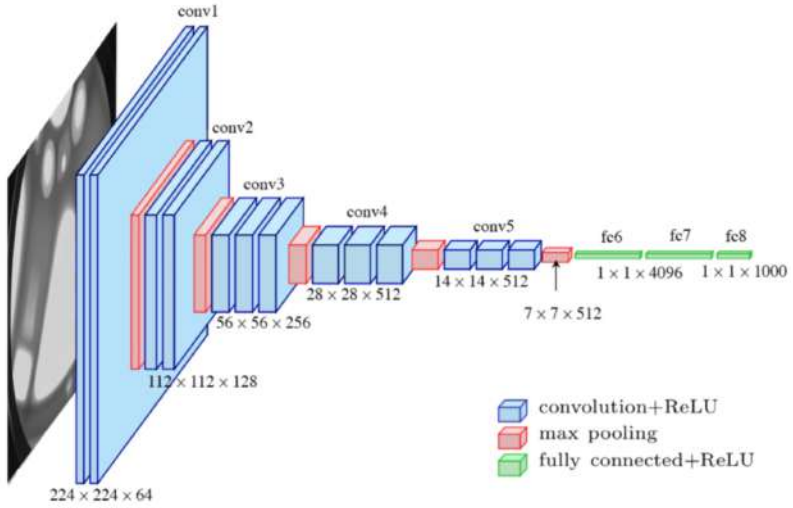


Fig. 1.2: Ejemplo de CNN con la arquitectura VGG19 [9].

1.2.1. Capas convolucionales

Una convolución es una operación matemática entre dos funciones que produce una tercera función [10]. En muchas disciplinas, una de las funciones representa una *señal* (imágenes en nuestro caso) y la otra un *filtro* o *kernel* (núcleo). En el caso de convolución entre funciones discretas, se desplaza el filtro (en la práctica es de soporte acotado) por la señal y le calcula promedios ponderados locales, transformándola en una nueva. La contraparte de la convolución en la corteza visual es la existencia de neuronas que responden únicamente a regiones específicas de la escena visual.

Mediante el uso de diversos filtros, la capas convolucionales (*convolutional layers*) de una CNN aplican convoluciones a la capa anterior, obteniendo nuevas representaciones de la misma. En el caso de una imagen, estos filtros extraen características relevantes (*features*) como la presencia de líneas, bordes, sombras, entre otros (Fig. 1.3).

Dado que una imagen tiene varios canales, el filtro que se le aplica tendrá 3 dimensiones: ancho y alto de la imagen, y el número de canales de la misma. Esta convolución da como

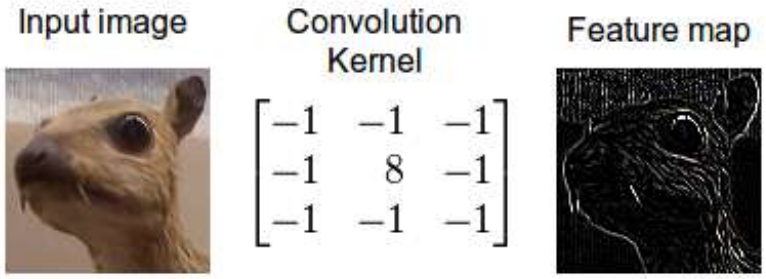


Fig. 1.3: Extracción de bordes mediante un filtro. Imagen obtenida de [11].

resultado un nuevo canal, pudiendo usarse varios filtros para generar varios canales, que juntos forman una nueva imagen. En arquitecturas como VGG19 (Fig. 1.2), el número de canales en las capas de convolución se duplica entre una y otra.

Por último, en todas las capas convolucionales se utiliza la función de activación ReLU (*Rectified Linear Unit*) para agregar no linealidad a la red y mejorar su capacidad de aprendizaje.

1.2.2. Capas de pooling

Las capas de pooling se encuentran usualmente luego de cada capa de convolución. Su función es realizar un submuestreo de dichas capas, reduciendo el alto y ancho de la imagen, para poder capturar aquellas caracterísiticas que sean invariantes antes pequeñas transformaciones, como una traslación o rotación, lo que puede ser útil para el reconocimiento de objetos. También ayuda a reducir la complejidad computacional y controlar el sobreajuste en el modelo. La contraparte del pooling en la corteza visual es la agregación de la respuesta de múltiples neuronas como estímulo para una neurona capaz de responder ante patrones más complejos en la escena visual.

Típicamente se implementa un *max-pooling*, que consiste en trasladar, con o sin superposición, una grilla (2x2, por ejemplo) a través de cada canal de la imagen y quedarse únicamente con el máximo valor. Un ejemplo se ilustra en la Figura 1.4.

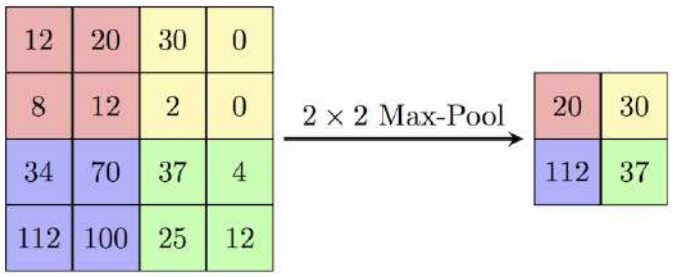


Fig. 1.4: Max-Pooling con una grilla 2x2 sin superposición. Imagen obtenida de [12].

1.2.3. Capas densas

Las capas densas se ubican al final de la CNN y cumplen el rol de condensar la información aún más de lo que se hizo en las capas previas. En las CNNs de clasificación de

imágenes, son las encargadas de etiquetar, mediante la función logit, las imágenes que se le pasen.

La capa densa, también conocida como capa totalmente conectada o *fully connected layer* en inglés, se ubican al final de una CNN y tiene un rol importante en el procesamiento y la clasificación de las características extraídas. Estas capas se llaman "densas"porque cada neurona en la capa está conectada a todas las neuronas en la capa anterior.

A diferencia de las capas convolucionales y de pooling, que se encargan de extraer características visuales en forma de mapas de características (*feature maps*), el propósito de las capas densas es aprender y combinar características de alto nivel para realizar la clasificación o regresión.

La cantidad de neuronas en la capa densa generalmente depende del número de clases o de la complejidad del problema que se está abordando. Por ejemplo, en un problema de clasificación con 10 clases, es común tener 10 neuronas en la capa densa final, donde cada neurona representa una clase diferente y se encarga de calcular la probabilidad de pertenencia a esa clase. La contraparte de estas unidades en la corteza visual es la existencia de neuronas capaces de responder ante la presencia de objetos específicos en la escena visual [3].

2. ACTIVACIONES BIOLÓGICAS Y ARTIFICIALES

En este capítulo presentamos y exploramos los datos que utilizaremos para investigar la correspondencia que existe entre el procesamiento de información en redes neuronales convolucionales y en el cerebro humano.

2.1. Dataset de imágenes

Utilizamos THINGS [13], un dataset de más de 26 mil imágenes específicamente diseñado para estudiar la visión humana. Cada imagen está manualmente etiquetada en uno de los 1.854 conceptos disponibles y cada uno de estos contiene un mínimo de 12 imágenes diferentes. Se busca que dentro de cada concepto las imágenes tengan variabilidad, que no se parezcan demasiado entre sí y que sean fácilmente identificables. Ejemplos de imágenes en este dataset se muestran en Fig. 2.1.

Fig. 2.1: Ejemplo de dos conceptos del dataset THINGS [13].

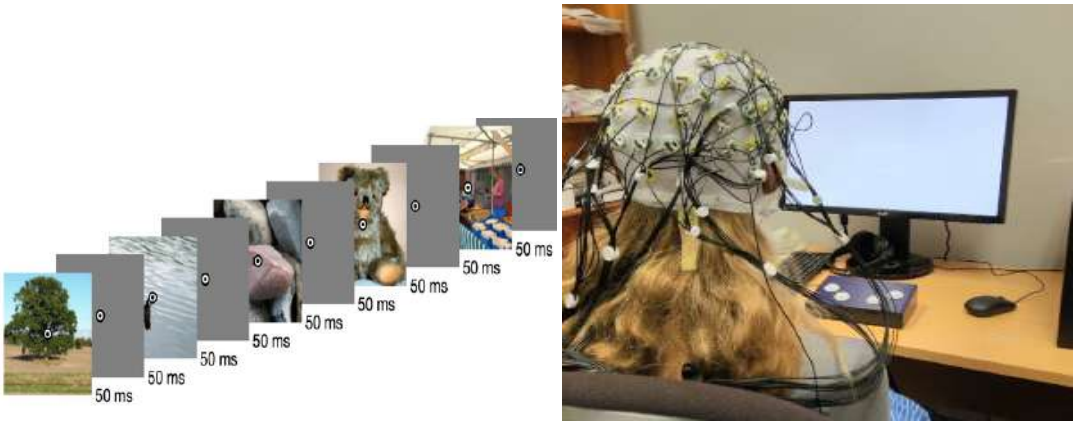
2.2. Activaciones electrofisiológicas

Grootswagers *et al.* [14] construyeron un dataset, THINGS-EEG, en base a las respuestas electroencefalográficas (EEGs) de 50 individuos con el objetivo de ser usado en investigaciones del sistema visual humano.

Cada participante se sienta frente a una pantalla y se le presentan secuencialmente 22.248 imágenes del dataset THINGS [13] etiquetadas en 1.854 conceptos. Cada imagen permanece 50 ms, seguida de una pantalla en blanco por otros 50 ms (Fig. 2.2). Es importante notar que dada esta frecuencia de presentación de imágenes, los participantes no reportan la experiencia subjetiva (consciencia) de haber percibido las imágenes.

Las EEGs se obtienen mediante un arreglo de 64 electrodos en la cabeza que muestrean a 1.000 Hz, que luego son filtradas para quedarse principalmente con las frecuencias entre 0,1 Hz y 100 Hz. Además, se las submuestrea a 250 Hz.

También realizaron el mismo procedimiento pero con un conjunto de 200 imágenes de conceptos diferentes, pero dentro de los 1.854. Estas provienen del mismo dataset [13] pero ninguna es de las 22.248 iniciales. Grootswagers *et al.* se valieron de este conjunto de EEGs más pequeño para realizar validaciones iniciales de lo que desarrollaron, pero nosotros lo utilizaremos únicamente para las tareas de predicción (ver capítulo 5).



(a) *(b)*

Fig. 2.2: Recortes de Fig. 1 de [14] *(a)* Ejemplo de sucesión de imágenes presentadas. *(b)* Setup del experimento.

Se excluyen del dataset las señales de un individuo debido a la mala calidad de estas, lo cual puede haber resultado de una muy alta impedancia de los electrodos de EEG. Por lo tanto, se tienen 49 registros EEG en total.

2.3. Activaciones de la CNN

Para obtener las activaciones en las capas sucesivas de una CNN partiendo de las imágenes, utilizamos una CNN con arquitectura VGG19 ([9], Fig. 1.2), la cual se encuentra preentrenada con el dataset ImageNet [15] de más de 3 millones de imágenes con sus conceptos correspondientes.

Trabajamos en Python 3.11.4, principalmente con las librerías NumPy y TensorFlow. Esta última provee, junto con Keras, un conjunto de herramientas para trabajar con redes neuronales, como es el caso de VGG19.

Al usar una imagen de entrada en la red se pueden obtener las activaciones en todas las capas intermedias, además del resultado de la clasificación a la salida de la capa densa final. En nuestro caso obtenemos las activaciones de las 5 capas de pooling (*pool1* a *pool5*) y de la segunda capa densa (*fc2*).

Como mencionamos en el Capítulo 1.2, entre una capa de pooling y la siguiente se reduce a la mitad el ancho y alto de las activaciones, y se duplica el número de canales. Al comienzo la imagen tiene dimensiones $224 \times 224 \times 3$ (ancho $\times$ alto $\times$ canales). La imagen pasa por el primer conjunto de capas convolucionales que convierten los 3 canales en 64 nuevos, y luego por *pool1* que reduce el ancho y el alto, dejando una activación o *mapa de features* de dimensiones $112 \times 112 \times 64$. A la salida de *pool5*, la activación tiene dimensiones $14 \times 14 \times 512$ (Fig. 2.3).

De acuerdo con estudios previos, en las primeras capas se interpreta que las activaciones representan aspectos de bajo nivel de la imagen (ej.: bordes, color, textura), mientras que en las últimas capas representan contenido semántico (ej.: árbol, pelota, silla) [5].

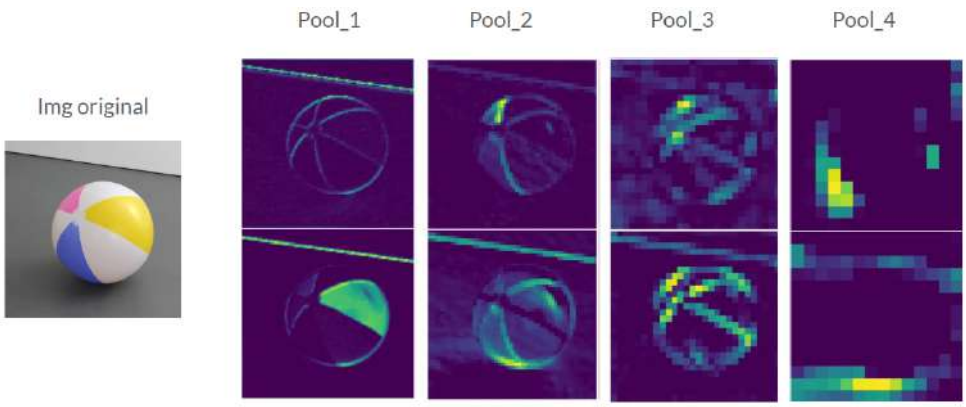


Fig. 2.3: Pares de canales de activaciones en *pool1* a *pool4* para una imagen dada.

3. ANÁLISIS DE SIMILITUD REPRESENTACIONAL - MATRICES DE DISIMILITUD

3.1. Descripción

El análisis de similitud representacional (RSA, por sus siglas en inglés) es una técnica utilizada en el ámbito de las ciencias cognitivas para investigar cómo se representan y procesan los estímulos en el cerebro [16]. RSA se basa en la premisa de que la información se codifica en patrones de actividad neuronal y que la similitud entre los patrones de actividad refleja la similitud entre los estímulos. Asimismo, si se tienen dos representaciones de la misma información en sistemas disímiles (por ejemplo, en el cerebro humano y en una red neuronal artificial), RSA puede utilizarse para comparar dichas representaciones, estableciendo una correspondencia entre ambas en el caso en que la similitud de las representaciones obtenidas en ambos sistemas se encuentre correlacionada sobre un determinado conjunto de estímulos.

Para aplicar el procedimiento de RSA se recopilan datos de imágenes, sonidos u otros estímulos presentados a los participantes mientras se registran sus respuestas neuronales, como la actividad cerebral medida mediante EEG. Luego, se calcula una matriz de similitud (o disimilitud) que compara cómo se correlacionan los patrones de actividad neuronal resultantes ante los diferentes estímulos. Esta matriz de similitud se somete a un análisis estadístico para revelar patrones y estructuras subyacentes a la representación neuronal. RSA permite identificar qué características o atributos de los estímulos son importantes para el cerebro y cómo se organizan esos atributos en la representación neuronal. También se puede utilizar para comparar la similitud entre diferentes regiones cerebrales y examinar la consistencia de la representación neuronal en diferentes contextos o tareas. El procedimiento se ilustra en la Fig. 3.1.

3.2. Motivación

Dado que no podemos comparar directamente las señales electrofisiológicas obtenidas mediante EEG (series de tiempo) con las activaciones de la CNN (mapa de características en forma de imágenes), necesitamos transformar ambas representaciones en una estructura en común que nos permita vincularlas.

Haciendo uso del método RSA, construimos matrices de disimilitud representacional (RDM por sus siglas en inglés) de los conceptos que representan las señales EEG y las activaciones, para poder establecer una correspondencia cuantitativa entre ambas.

3.3. RDM - EEG

La idea principal llevada a cabo por Grootswagers *et al.* [14] para armar las RDMs en base a las activaciones EEGs, consiste en definir la disimilitud como la capacidad de un clasificador lineal para diferenciar las activaciones neuronales asociadas a un concepto con respecto a las de otro.

Dado un individuo y un tiempo determinado, calculan la RDM de los 1.854 conceptos. Para cada par de conceptos, realizan una validación cruzada de tipo *leave-one-out* (12

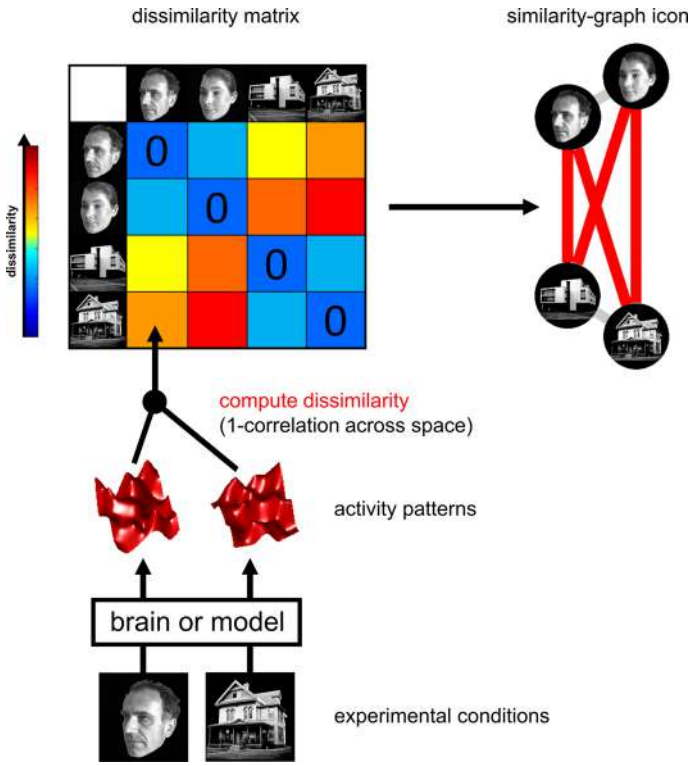


Fig. 3.1: Ejemplo ilustrativo de la metodología RSA. Ante un conjunto determinado de estímulos (en este caso, dos caras y dos casas), se computan las disimilitudes entre todos los pares. La disimilitud puede computarse de distintas formas, por ejemplo, entre las imágenes originales, pero también entre la respuesta cerebral evocada por las mismas. La correlación entre dos matrices de disimilitud sugiere que la información representada en ambas instancias puede ponerse en correspondencia [16].

respuestas EEGs por concepto) y en cada paso entrenan un clasificador lineal, donde la variables predictoras son las EEGs y las variables a predecir son los conceptos. La exactitud o *accuracy* promedio de los clasificadores es usada como medida de disimilitud entre el par de conceptos. De esta forma, un mejor desempeño del clasificador indica que las imágenes son más disímiles entre sí, mientras que imágenes muy similares resultarán en desempeños cercanos o iguales al nivel chance.

El rango de tiempo utilizado empieza a los 100 ms antes de presentarle la imagen al individuo hasta los 1000 ms después de dicho evento. Estos 1.100 ms, estando digitalizados a una frecuencia de muestreo de 250 Hz, se corresponden a 275 muestras temporales. Por lo tanto, se consiguen $1.854 \times 1.854 \times 275 \times 49$ (concepto $\times$ concepto $\times$ tiempo $\times$ individuo) RDMs (una por cada muestra en el tiempo).

3.4. RDM - CNN

Queremos calcular la RDM de las activaciones para cada capa de interés de la CNN. Sean $x_1, x_2 \in R^p$ un par de activaciones provenientes de dos imágenes a las cuales se les aplana las dimensiones en un vector (ej.: una activación de *pool1* pasa de tener dimensiones $112 \times 112 \times 64$ a $p = 802.816$), definimos la disimilitud entre ambas como su descorrelación de Spearman:

$$1 - Spearman(x_1, x_2) \in [0, 2]$$

donde $Spearman(x, y) \in [-1, 1]$ es la correlación de Spearman entre x e y.

Elegimos usar esta correlación ya que permite comparar la monotonía entre ambas activaciones, sin asumir necesariamente una relación lineal entre ambas variables. Por ejemplo, si ambas presentan valores altos o valores bajos en las mismas posiciones, la correlación será alta.

Para construir la RDM de cada capa, idealmente nos gustaría usar todas las 22.248 imágenes a la vez. Calcularíamos entonces la disimilitud entre dos conceptos como el promedio de descorrelaciones entre cada una de las 12 imágenes de un concepto con las 12 de otro. Lamentablemente surgen dos problemas de índole computacional:

- no entran simultáneamente todas las imágenes en la memoria RAM de la computadora que usamos, la cual tiene 64 Gb;
- intentar hacerlo por lotes resulta en tiempos excesivos de cómputo (varios días).

Para sortear estos obstáculos, optamos por construir 12 RDMs, cada una obtenida usando una sola imagen por concepto en vez de 12, sin repetir imágenes entre conceptos. La factibilidad de esta estrategia se sustenta en el hecho de que las distintas imágenes de un mismo concepto son similares entre sí y en consecuencia existe redundancia en esta información.

Un ejemplo de matriz de RDM computada de las activaciones de la CNN se muestra en la Fig. 3.2.

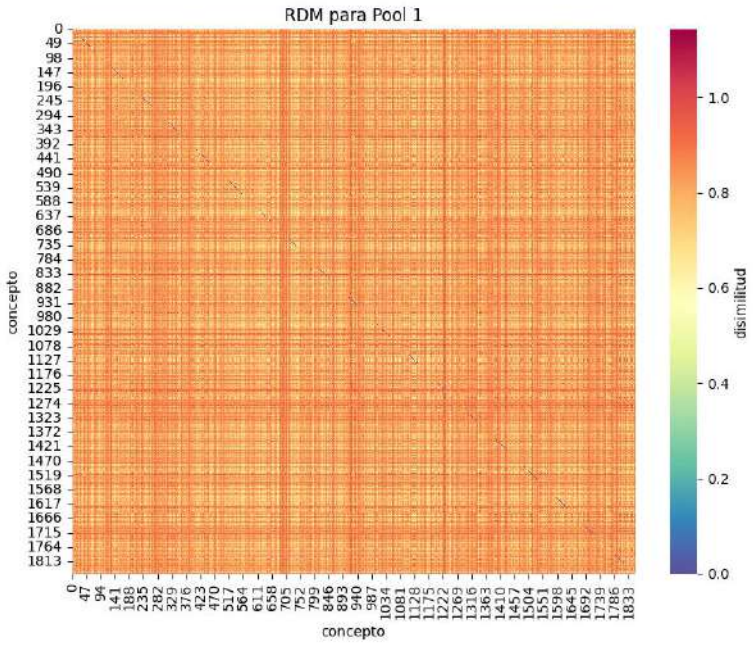
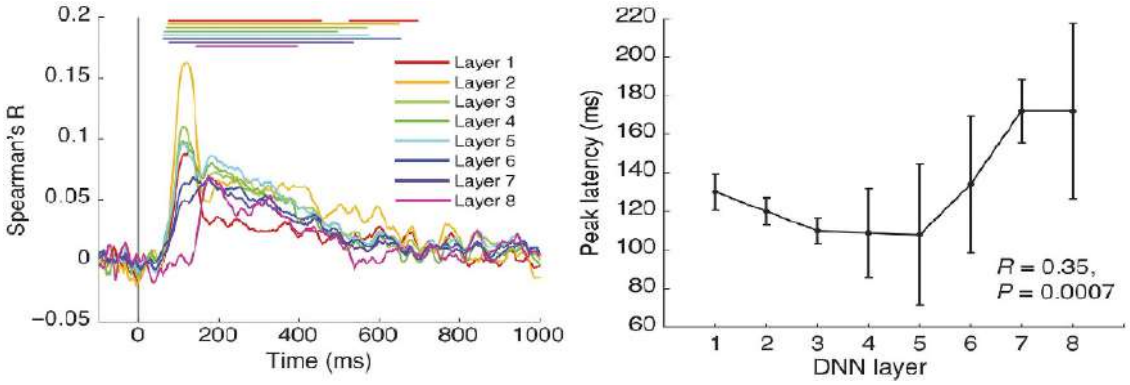


Fig. 3.2: Una RDM para *pool*1.

4. CORRESPONDENCIA ENTRE RESPUESTA DE EEG Y ACTIVACIONES EN CNN

Una vez obtenidas las RDMs de la CNN y las señales EEG, podemos compararlas y analizar si existe una correspondencia entre ambas representaciones. Específicamente, queremos observar las correspondencias a través del tiempo, diferenciando por cada capa de interés de VGG19.

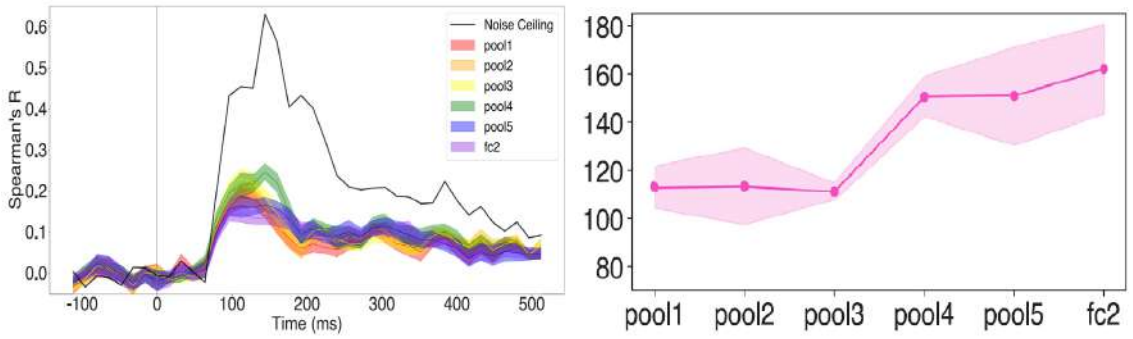
También nos interesa ver cuál es el tiempo de correspondencia máximo para cada capa. Cichy *et al.* [17] y Güçlü *et al.* [18] sugieren que las primeras capas de una CNN se corresponden mayormente con las etapas tempranas del sistema visual humano, mientras que las capas avanzadas se corresponden con etapas superiores del sistema visual. Por lo tanto, es de esperar que exista una relación monótona entre el número de capa y el tiempo de máxima similitud entre las RDMs de EEG y las RDMs correspondientes a cada capa. En vez de analizar la correspondencia utilizando señales EEG, Cichy y colegas lo hacen con señales MEG (magnetoencefalograma) tomadas de 15 individuos al mostrarles 118 imágenes de objetos. Esta técnica permite obtener señales de mayor calidad ya que presenta una mejor relación señal-ruido. También utilizan una CNN con un mayor número de capas. Los resultados obtenidos son consistentes con la hipótesis, tal como se muestra en la Fig 4.1.



(a) *(b)*

Fig. 4.1: Recortes de Fig. 3 de Cichy *et al.* [17]. *(a)* Correspondencia temporal diferenciada por capa de la CNN. *(b)* Tiempos de correspondencia máxima para cada capa de la CNN.

Un trabajo similar fue realizado por Kong *et al.* [19], utilizando una CNN de arquitectura VGG19 pero con un dataset de señales EEG distinto al nuestro. Mostraron que utilizando una medida de similitud apropiada se observa una tendencia positiva entre los tiempos de correspondencia máxima y el número de capa, tal como es esperable de acuerdo a los resultados de Cichy y Güçlü [17, 18]. También calcularon, para cada tiempo, la combinación lineal de activaciones de CNN que mejor se correlaciona con las EEGs. Con esta optimización obtuvieron valores de correspondencia más altos. Estos resultados se muestran en la Fig. 4.2.



(a) *(b)*

Fig. 4.2: Recortes de Fig. 6 (izquierda) y Fig. 8 (derecha) de Kong *et al.* [19]. *(a)* Correspondencia temporal diferenciada por capa de VGG19. *(b)* Tiempos de correspondencia máxima para cada capa de VGG19.

4.1. Metodología

Sea $RDM_{EEG}(i,t)$ la RDM de un individuo i a tiempo t ($i \in [1..49]$, $t \in [1..275]$), y sea $RDM_{CNN}(j,c)$ la RDM asociada al conjunto de imágenes j ($j \in [1..12]$) y la capa de interés c ($c \in [1..6]$). La correspondencia entre $RDM_{EEG}(i,t)$ y $RDM_{CNN}(j,c)$ viene dada por la correlación de Spearman entre los vectores obtenidos al aplanar la triangular superior de ambas matrices, donde luego promediamos a través de j. En total, tenemos $49 \times 275 \times 6$ (individuo $\times$ tiempo $\times$ capa) valores de correspondencia $C(i,t,c)$.

4.2. Resultados

4.2.1. Con todos los individuos

Para visualizar los resultados, en una primera instancia promediamos estas correspondencias a través de los individuos,

$$\frac{1}{49}\sum_{i=1}^{49} C(i,t,c)$$

obteniendo 275×6 valores. Además, estimamos el error de dicho promedio:

$$\frac{std_i\Big(C(i,t,c)\Big)}{\sqrt{49}}$$

donde $std_i(.)$ es la desviación estándar a través de i. Estos resultados se muestran en la Fig. 4.3.

Recordemos que cada imagen se muestra por 50 ms (intervalo [0, 50] ms según nuestro sistema de referencia), por lo que durante los 100 ms previos y los 950 ms que vienen después de que esta finalice aparecerán otras imágenes. A pesar de esto, al correlacionar con las activaciones de VGG19 de dicha imagen solo aumenta visiblemente la correlación en el rango [100; 300], observándose principalmente dos picos. Antes de los 100 ms y después de los 400 ms la correlación es prácticamente cero. Esto implica que, a pesar de que se exponen varias imágenes en el periodo evaluado ([-100; 1000] ms), solo tendrá incidencia

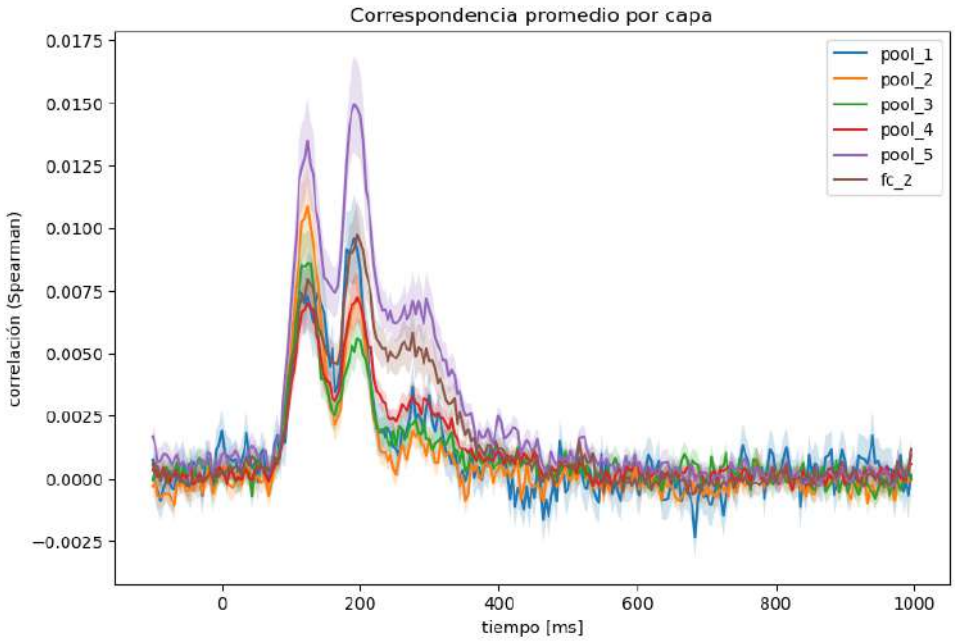


Fig. 4.3: Correspondencia promedio entre respuesta de EEG y capas de la red para los 49 individuos a lo largo del tiempo, diferenciada con colores por cada capa de VGG19.

la imagen usada para obtener las activaciones de VVG19 que son correlacionadas con las respuestas EEG. Esto es debido a que las imágenes que se presentan antes y después de una imagen dada pertenecen a distintos conceptos y por lo tanto no sesgan sistemáticamente la respuesta del EEG y la subsiguiente construcción de las RDMs en base al clasificador lineal.

Por otro lado, definimos el tiempo de correspondencia máximo de una capa c' como el promedio, a través de los individuos, de los tiempos donde su correspondencia es máxima.

$$\frac{1}{49}\sum_{i=1}^{49}\operatorname*{argmax}_{t} C(i,t,c')$$

No se observa (Fig. 4.4) una forma similar a la de Kong *et al.* [19] (Fig. 4.2) o Cichy *et al.* [17] (Fig. 4.1). En nuestro caso la tendencia, a medida que avanzamos de capa, es negativa en la mayoría de los casos. Además, en la figura de correspondencia temporal (Fig. 4.3) puede apreciarse claramente que los picos están entre los 100 y los 200 ms, pero los promedios de tiempos máximos tienen valores más altos. Esta discrepancia podría aparecer debido a que hay individuos cuya correspondencia no presenta picos claros y cuyo máximo podría asignarse, en principio, a cualquier instante de tiempo, afectando el promedio. Esto puede suceder por motivos tales como pérdida de atención a los estímulos visuales, o incluso episodios de pérdida de vigilancia y sueño durante el experimento.

4.2.2. Resultados con individuos seleccionados

Haciendo una inspección visual, identificamos algunos individuos con mala o ruidosa correspondencia. Estos sujetos no presentan picos de correspondencia en el rango de 100 a 200 ms y sus valores son mucho más bajos que la correspondencia promedio (Fig. 4.3). El

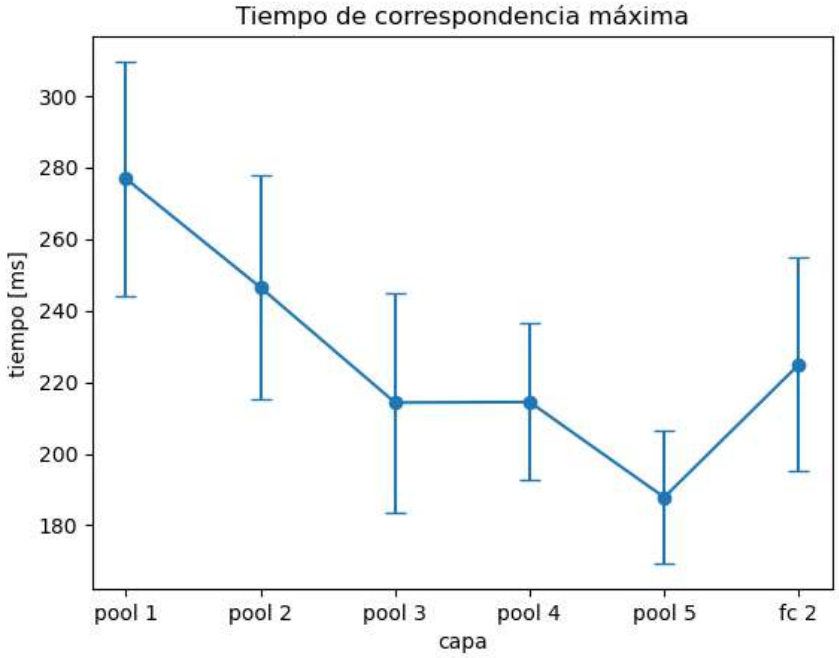


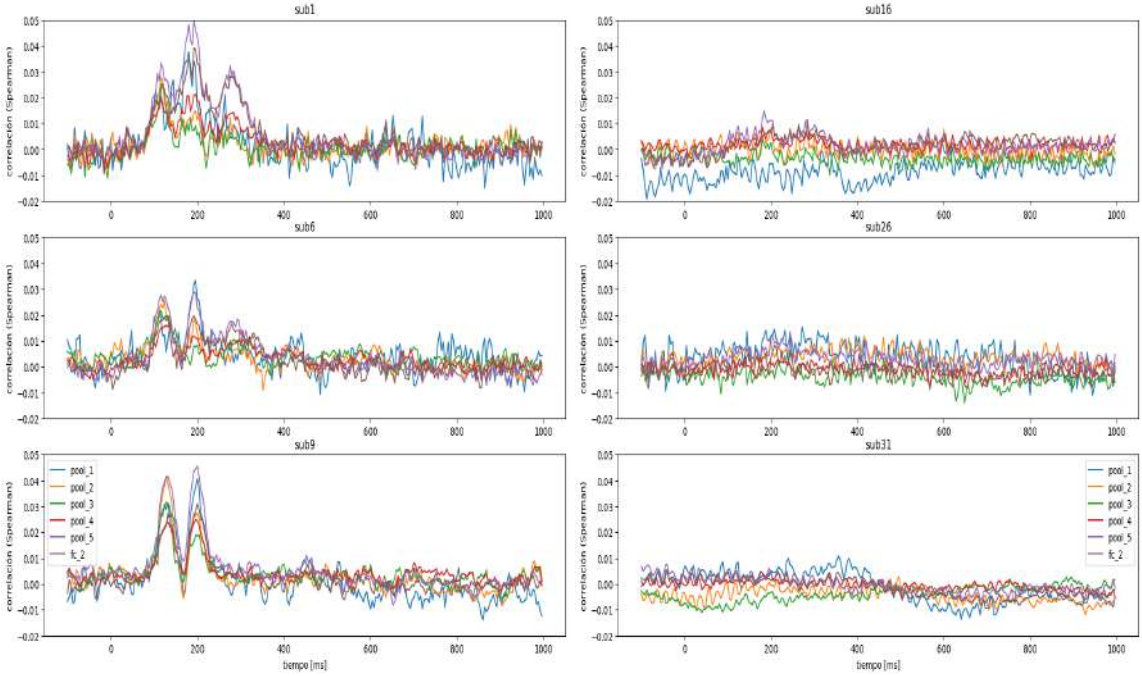
Fig. 4.4: Promedio entre individuos de los tiempos de correlación máxima.

criterio elegido para seleccionarlos automáticamente es que el máximo de correspondencia a través del tiempo y las capas sea menor a 0.02 (Fig. 4.5).

Luego de quitar estos casos, nos quedamos con un total de 21 individuos. Decidimos volver a calcular la correspondencia media y los tiempos de correspondencia máxima removiendo los individuos que no presentan una clara correspondencia. En la Fig. 4.6 se muestran los resultados de las correspondencias vs. tiempo para estos sujetos, y en la Fig. 4.7 el momento de máxima correspondencia para cada capa de la red.

Estos nuevos valores de correspondencia arrojan picos de mayor correlación. Además, aparece también una tendencia positiva entre los tiempos de correspondencia máxima de $pool3$ a $fc2$, tal como es de esperarse a partir de los trabajos previos mencionados anteriormente (aunque el tiempo para $pool1$ sigue estando arriba del resto).

Algo importante a observar en las figuras de correspondencia (Fig. 4.3 y 4.6) es la altura relativa entre los dos picos de una capa. En el caso de $pool2$ y $pool3$, el primer pico está mucho más arriba del segundo; para $pool4$ están prácticamente iguales y para $pool5$ y $fc2$ ya se observa que el segundo pico está arriba del primero. En estas última dos capas también ocurre que aparece un tercer pico, que en las anteriores no es muy notorio. Todo esto refuerza la hipótesis inicial de que, a excepción de la primera capa ($pool1$), existe una relación monótona creciente entre el número de capas y el tiempo de correlación máxima.



(a) Individuos $sub1, sub6, sub9$ con buena correspondencia.

(b) Individuos $sub16, sub26, sub31$ con mala correspondencia.

Fig. 4.5: Ejemplos de buenas correspondencias (izquierda) y malas correspondencias (derecha).

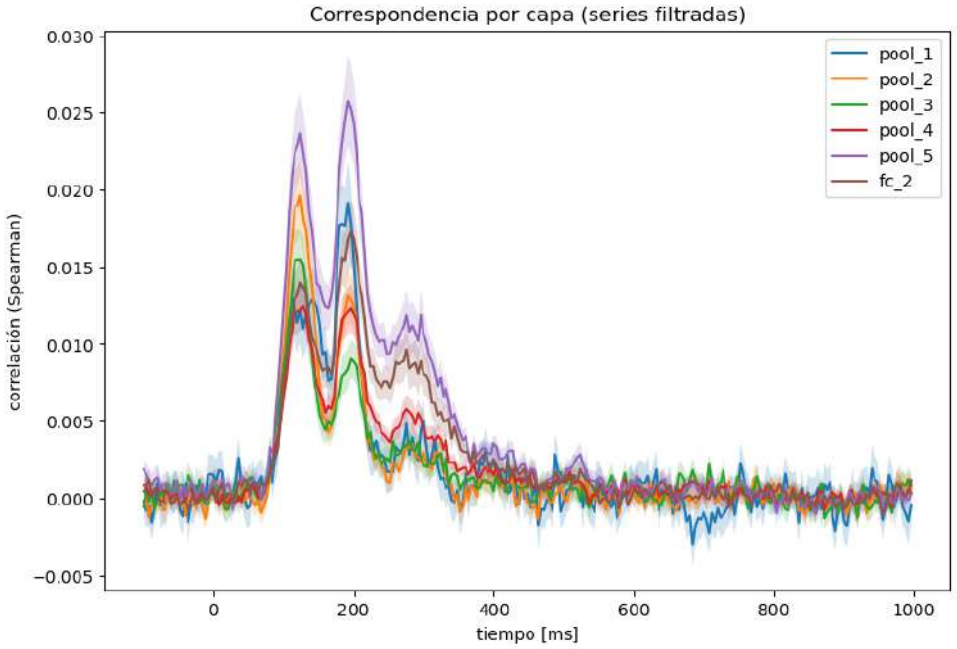


Fig. 4.6: Correspondencia promedio entre los 21 individuos que quedaron luego de eliminar las malas correspondencias, diferenciadas por capa de VGG19.

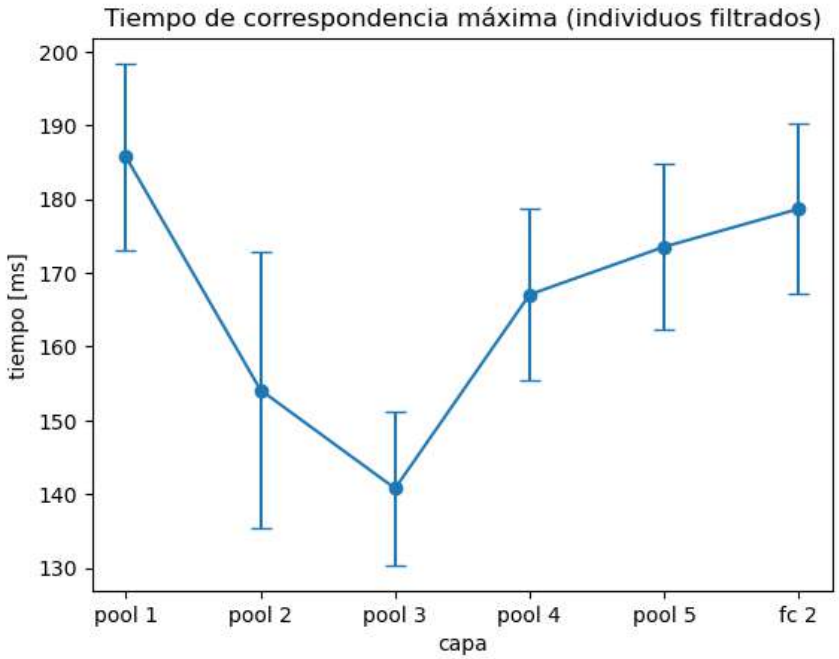


Fig. 4.7: Promedio entre los 21 individuos, que quedaron luego de eliminar las malas correspondencias, de los tiempos de correlación máxima.

5. PREDICCIÓN DE RESPUESTAS EEG A PARTIR DE ACTIVACIÓN EN CNN

En el Capítulo 4 analizamos la correspondencia calculando la correlación entre las respuestas EEG y las activaciones de VGG19, reproduciendo resultados publicados previamente por [19] y colegas. En este capítulo exploramos nuevas formas de analizar el vínculo entre ambas representaciones. Para ello, en vez de utilizar RSA, investigamos la posibilidad de predecir las señales EEG en base a las activaciones obtenidas en la CNN. En particular, investigamos si los momentos en los cuales dicha predicción es óptima se encuentran correlacionados con el número de capa de la red que fue usada para la predicción.

5.1. Metodología

Llamamos $A_c \in R^{n \times p_c}$ a la matriz resultante de poner como filas las n activaciones vectorizadas de longitud p_c, provenientes de la capa c ($c \in [1..6]$). Por otro lado, tenemos las respuestas EEG, $Y(i,t) \in R^n$, asociadas a las mismas n imágenes, de un individuo i a tiempo t ($i \in [1..49]$, $t \in [1..275]$).

Fijando c, i y t, queremos estimar qué tan bueno es el ajuste de un modelo lineal que tome las columnas y las filas de A_c como variables y observaciones, respectivamente, e intente predecir los valores de $Y(i,t)$. Con el fin de evitar evaluar el modelo con observaciones usadas para entrenarlo, realizamos una validación cruzada al estilo *k-folds* de a 5 bloques (k = 5). Para ello, se subdivide la matriz A_c y el vector $Y(i,t)$ a lo largo de las observaciones en k partes de tamaño lo más similar posible y en cada paso se reserva una parte como conjunto de testeo (20 %). El restante (80 %) se usa para entrenar el modelo. Como medida de correspondencia, tomamos el R^2 promedio obtenido en cada bloque, totalizando 6 × 275 × 49 (capa × tiempo × individuo) valores. Este procedimiento se ilustra en la Fig. 5.1.

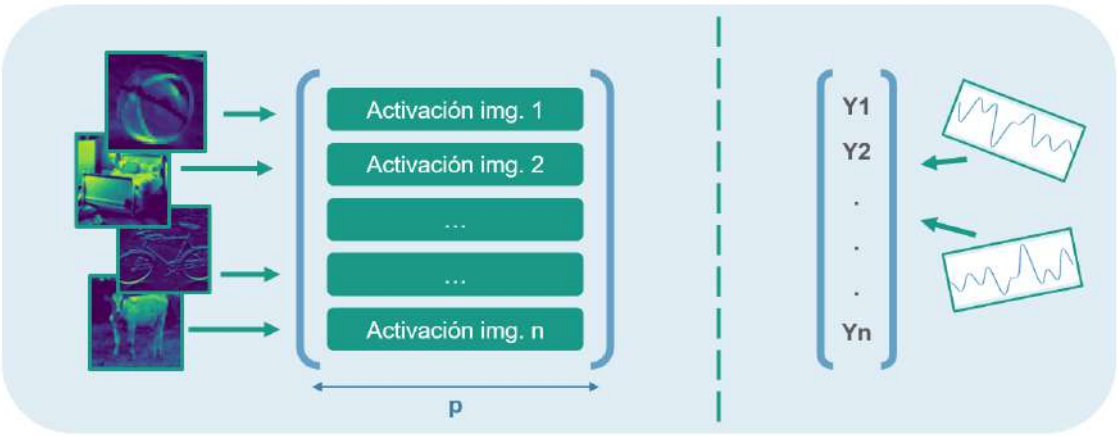


Fig. 5.1: Diagrama de la matriz de activaciones (izquierda) y el vector de respuestas EEG a las imágenes en un tiempo determinado (derecha).

Agregamos una regularización Ridge para evitar un sobreajuste del modelo lineal. Además, debido a que los valores de p_c son mucho más grandes que n (ej.: para *pool1* se tiene $p = 802.816$), los tiempos de cómputo serían muy altos. Decidimos entonces realizar una reducción del espacio de *features* aplicando PCA (análisis de componentes principales) de cada matriz A_c y quedarnos con las primeras $\lfloor 0{,}8n \rfloor$ componentes principales. Obtenemos nuevas matrices $\tilde{A}_c \in R^{n \times \lfloor 0{,}8n \rfloor}$ en base a los *scores* de cada componente y realizamos el ajuste lineal con ellas. Al ponderar a n con 0,8 nos garantizamos que, al momento de la validación cruzada, el número de variables no exceda al de observaciones.

Al igual que en el Capítulo 4, obtenemos los tiempos promedios de correspondencia (R^2) máxima y volvemos a analizar si existe una relación creciente entre los tiempos y el número de capa.

5.2. Resultados

Utilizamos el conjunto de 200 imágenes y las 49 × 200 (individuo × imágenes) EEGs asociadas, descritas en el Capítulo 2. La varianza total explicada por las primeras $0,8n = 160$ componentes principales de las matrices de activaciones A_c está en el orden del 80 %.

Promediamos los 6 × 275 × 49 (capa × tiempo × individuo) valores de R^2 a lo largo de los individuos y estimamos el error del promedio como la desviación estándar entre los individuos dividido la raíz cuadrada de la cantidad de los mismos (mismo procedimiento que realizamos en el Capítulo 4 para la correspondencia). Los resultados se muestran en la Fig. 5.2.

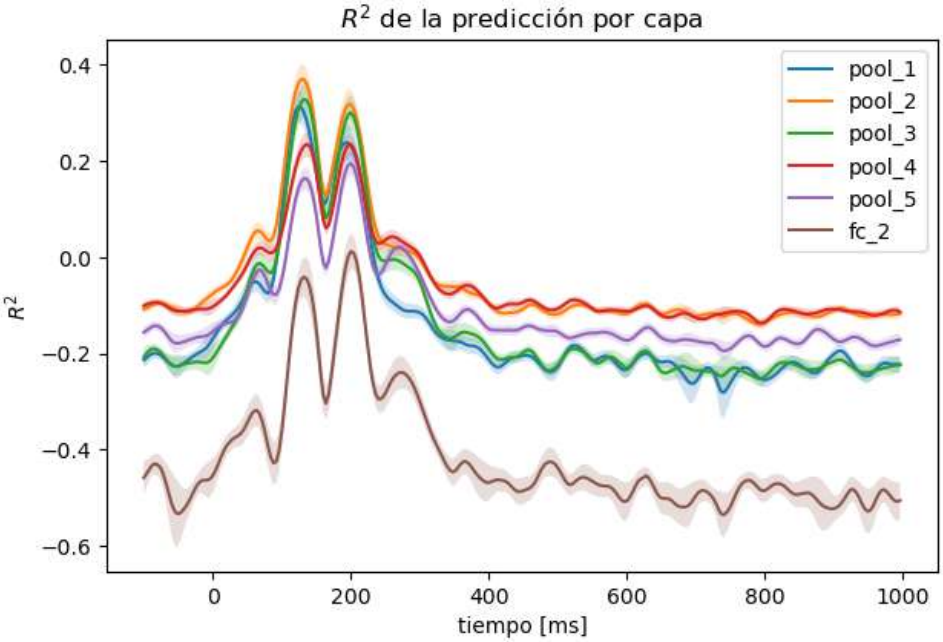


Fig. 5.2: R^2 promedio de cada predicción a lo largo del tiempo y diferenciado por capa.

Si bien los valores de R^2 obtenidos no son muy buenos, nuestro objetivo no es evaluar el desempeño del clasificador sino en estudiar como el mismo cambia a lo largo del tiempo; en particular, en qué instante de tiempo dicho desempeño es óptimo. Notamos que aparecen valores más altos entre los 100 y 300 ms, al igual que en la correspondencia del Capítulo

4, indicando que se logró predecir mejor la respuesta de EEG en dicho intervalo, y por lo tanto evidenciando mediante un método distinto que es en dicho intervalo de tiempos donde se encuentra la mayor cantidad de información de la respuesta ante los estímulos visuales.

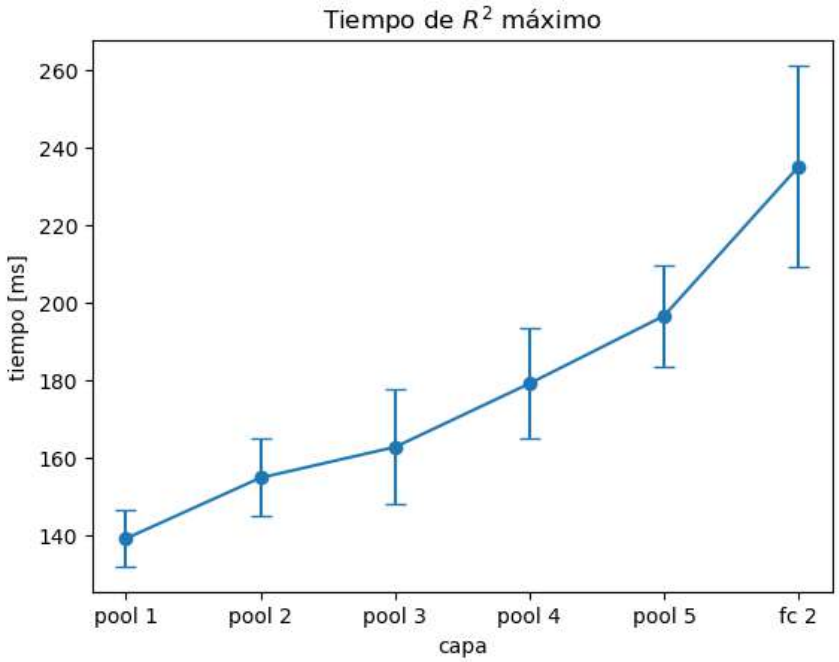


Fig. 5.3: Tiempo de R^2 máximo, diferenciado por capa.

En el caso de los tiempos de R^2 máximos es donde obtenemos resultados que apoyan más fuertemente la hipótesis de correspondencia entre respuesta de EEG y activaciones de CNN. A diferencia de los resultados obtenidos en la correspondencia (Fig. 4.4 y 4.7), en la Fig. 5.3 apreciamos una fuerte relación monótona creciente entre los tiempos y el número de capa.

6. CONCLUSIONES

6.1. Resultados obtenidos

En el Capítulo 4 cuantificamos la correspondencia entre las EEGs y la CNN calculando la correlación de Spearman entre las RDMs de las EEGs de todos los individuos y las RDMs de las activaciones de la CNN. No obtuvimos resultados que apoyen fuertemente la hipótesis de una relación monótona creciente entre el número de capas y el tiempo de correspondencia máximo (Fig. 4.4). De todos modos, pudimos observar que, moviéndonos de *pool2* a *fc2*, la altura de los dos picos principales pasaba de ser mayor para el primero a ser mayor para el segundo, indicando una relación entre los tiempos de correspondencia alta y el número de capa. Al quitar a los individuos que presentaban correspondencias sin picos claros pudimos acercarnos más a la hipótesis, particularmente entre las capas *pool3* y *fc2*. Además obtuvimos mayores valores de correspondencia máxima (Fig. 4.6) y la curva en la figura 4.7 presentó una relación creciente a partir de *pool3*. Aún así, los valores de correspondencia de la capa *pool1* se mantuvieron altos con respecto a *pool2* o *pool3*, con el segundo pico arriba del primero.

Más adelante, en el Capítulo 5 investigamos qué tan bien podemos predecir las EEGs tiempo a tiempo usando modelos lineales que tomen las activaciones de la CNN como variables predictoras. Tomamos el R^2 de cada predicción como medida de la calidad del ajuste. A pesar de que los valores de R^2 a lo largo del tiempo no evidencian una alta capacidad de predicción, las formas de dichas curvas (Fig. 5.2) se asemejan a las figuras 4.3 y 4.6 del Capítulo 4. Además, en este caso observamos una tendencia positiva entre el número de capa y el tiempo de R^2 máximo.

6.2. Comparación con otros estudios

Uno de los artículos que usamos de base para nuestros estudio fue el de Cichy *et al.* [17], quienes investigaron la correspondencia utilizando una red neuronal más profunda que VGG19 y también midiendo las respuestas neuronales mediante MEG en vez de EEG. Generalmente, la técnica de MEG es capaz de registrar señales con una mejor relación entre señal y ruido, proporcionando una potencial mayor robustez a la hora de calcular las RDMs. Aún así, los tiempos de correspondencia máxima ilustrados en la figura 4.1b se asemejan a los nuestros, a excepción del valor en la capa *pool1*.

Otro artículo considerado fue el de Kong *et al.* [19]. En este estudio, los autores se propusieron encontrar para cada tiempo la mejor combinación lineal de activaciones que maximiza la correlación con las matrices de disimilitud de EEG. Naturalmente, este proceso de optimización resulta en valores de correspondencia mucho mayores al nuestro, el cual no fue optimizado de esta manera. No obstante, es importante remarcar que los coeficientes de la combinación lineal buscada por Kong *et al.* son distintos para cada muestra temporal. Consideramos cuestionable el hecho de optimizar para cada tiempo distinto en vez de buscar una única combinación lineal que sirva para todos los tiempos, dado que es difícil intepretar biológicamente la dependencia en el tiempo de la combinación lineal óptima. Por otra parte, nuestra metodología encuadra mejor con la idea de buscar un vínculo entre ambas representaciones que sea agnóstico respecto al tiempo.

6.3. Limitaciones de EEG

Si bien la técnica de EEG permite evaluar la correspondencia en el tiempo, no permite evaluarla anatómicamente. Técnicas como fMRI (imagen por resonancia magnética funcional), de mejor resolución espacial, permiten investigar si la corteza visual primaria, secundaria, etc., están en correspondencia jerárquica con las capas sucesivas de la red. Investigaciones como las de Xu y Vaziri-Pashkam [20] exploran esta opción en profundidad.

Otra limitación de la EEG es que las señales del cerebro capturadas pueden ser fácilmente afectadas por movimientos musculares de la cara por parte del participante, como puede ser un pestañeo. Esto conlleva a que se le pida al participante reducir al mínimo todos sus movimientos faciales, lo cual se dificulta en sesiones largas. Estos artefactos, sumados a la posible pérdida de concentración y vigilancia a lo largo del tiempo, pueden explicar la necesidad de remover una cantidad considerable de sujetos del análisis.

6.4. Limitaciones de las CNN

Las CNNs están inspiradas en la forma en que funciona el sistema visual humano pero no son una réplica exacta. Tienen principios similares de procesamiento visual jerárquico y extracción de características, pero esto no es suficiente para emular el sistema visual humano. Por ejemplo, la corteza visual primaria inicialmente representa en su actividad neuronal la presencia de segmentos orientados o una combinación de ellos [1, 2], mientras que en las primeras capas de las CNNs tienden a representar principalmente texturas y bordes. A su vez, las CNNs tienen una arquitectura *feedfoward*. Es decir, las activaciones siguen un orden secuencial sin haber componentes o conexiones recurrentes, las cuales sí están presentes en la corteza cerebral. Las CNNs como VGG19 tampoco tienen una noción completa del contexto de una imagen que sirva para clasificarla mejor.

6.5. Trabajo a futuro

Como posibles investigaciones futuras, proponemos aplicar las mismas metodologías, pero con otros datasets de imágenes, otros tipos de técnicas de captura de señales cerebrales (MEG, fMRI, etc.) y otras arquitecturas de CNN con más o menos capas que VGG19, comparándolas entre sí, usando la correlación entre número de capa y máxima correspondencia como indicador de similitud con el procesamiento de información en el cerebro.

Otra vía de investigación sería cambiar el diseño experimental de presentación de imágenes. En el experimento de Grootswagers *et al.* [14] los individuos no alcanzan la percepción consciente de las imágenes debido a los cortos tiempos de exposición (50 ms) sumados a la rápida aparición de una nueva imagen en la pantalla. Si bien la percepción consciente no es necesaria para que un estímulo visual se represente en la corteza cerebral, es posible que la robustez de dicha representación sea mayor cuando ocurre de manera consciente. Por lo tanto, podríamos realizar un nuevo experimento en el que se aumente el tiempo de exposición de cada imagen a una duración que habilite la percepción consciente por parte de los individuos. En ese caso, podríamos esperar resultados más claros debido a una mejor relación entre señal y ruido en la respuesta medida con EEG.

Bibliografía

[1] David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

[2] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[3] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

[4] Kunihiko Fukushima. Neocognitron. *Scholarpedia*, 2(1):1717, 2007.

[5] Neena Aloysius and M Geetha. A review on deep convolutional neural networks. In *2017 international conference on communication and signal processing (ICCSP)*, pages 0588–0592. IEEE, 2017.

[6] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Learning representations by back-propagating errors. *Nature*, 323, 1986.

[7] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, Howard R. E., W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. NIPS'89, page 396–404. MIT Press, 1989.

[8] CS231n: Deep Learning for Computer Vision - Stanford - Spring 2023. `https://cs231n.github.io/convolutional-networks/`. Accessed: 2023-09-25.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[10] Convolution - Wikipedia. `https://en.wikipedia.org/wiki/Convolution`. Accessed: 2023-09-25.

[11] Convolution - Nvidia. `https://developer.nvidia.com/discover/convolution`. Accessed: 2023-09-25.

[12] MaxPooling - Computer Science Wiki. `https://computersciencewiki.org,%/index.php/File:MaxpoolSample2.png`. Accessed: 2023-09-26.

[13] M. N. Hebart, Dickter A. H., A. Kidder, W. Y. Kwow, A. Corrieveau, C. Van Wicklin, and C. I. Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 2019.

[14] T. Grootswagers, I. Zhou, A. K. Robinson, M. N. Hebart, and T. A. Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 2022.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. 2009.

[16] Nikolaus Kriegeskorte and Rogier A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013.

[17] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6, 2016.

[18] Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[19] Nathan C.L. Kong, Blair Kaneshiro, Daniel L.K. Yamins, and Anthony M. Norcia. Time-resolved correspondences between deep neural network layers and eeg measurements in object processing. *Vision Research*, 172:27–45, 2020.

[20] Y. Xu and M. Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, 2021.