

TP Final

Grupo “Marcos & Marcos”

16.82 - Neurociencia Computacional

Primer cuatrimestre 2024



Integrantes:

- Marcos Casiraghi (62003) - mcasiraghi@itba.edu.ar
- Marcos Gronda (62067) - mgronda@itba.edu.ar

Profesores:

- Dante Kienigiel
- Facundo Emina

Indice

Indice	1
1. Introducción	2
2. Modelo	4
2.1 Layer-wise Relevance Propagation (LRP)	4
2.2 Relevancia de conexiones sinápticas	5
3. Resultados	6
3.1 Análisis de PCA	6
3.2 Análisis de LRP	7
4. Conclusión	11
5. Anexo	12
5.1 LRP de conexiones sinápticas: método relevancia promediada	12
5.2 LRP de conexiones sinápticas: método relevancia decadente	13
5.3 LRP de valores de entrada de 4 clases distintas	14

1. Introducción

Como ya se sabe, las redes neuronales probaron ser una gran herramienta con muchas aplicaciones como la clasificación de imágenes o reconocimiento del habla pero puede resultar complicado entender cómo es que realmente funciona por dentro. Muchas veces a las redes neuronales se las trata como una “caja negra” que recibe un input y devuelve una salida, donde el proceso de por medio se lo da por sentado. Las personas que utilizan las redes neuronales tal vez no saben que está haciendo o qué patrones está levantando del input que se le pasó. Una posible solución para este problema se conoce como Explainable AI, un término para los métodos que se utilizan para intentar hacer que una red neuronal tenga mayor sentido.

El objetivo de este trabajo es explorar este concepto de Explainable AI para poder observar y analizar qué es lo que aprende un perceptrón multicapa y de esta forma visualizar cómo es que la red aprende y selecciona para poder ver que hay dentro de esta “caja negra”.

Para la realización de este proyecto se trabajó con un dataset de respuesta emocional hacia ciertos videojuegos, emocionalmente clasificados con las siguientes descripciones: aburrido, tranquilo, de horror o divertido. Los 4 juegos son respectivamente: Train Sim World, Unravel, Slender - The Arrival y Goat Simulator. Los 28 participantes jugaron estos videojuegos por 5 minutos cada uno, sumando un total de 20 minutos jugados y se registró su actividad neuronal EEG con el dispositivo Emotiv Epoc+ de 14 canales.

En primer lugar es necesario procesar los datos EEG, por lo que se toman los datos de los canales: AF3, AF4, F3, F4, F7, F8, FC5, FC6, O1 y O2. En los 5 minutos donde cada participante jugó a cada videojuego, se registraron los valores cada 7.84 ms para obtener 38265 lecturas de todos los canales. A continuación se muestra una visualización de donde se encuentran los receptores para estos canales:

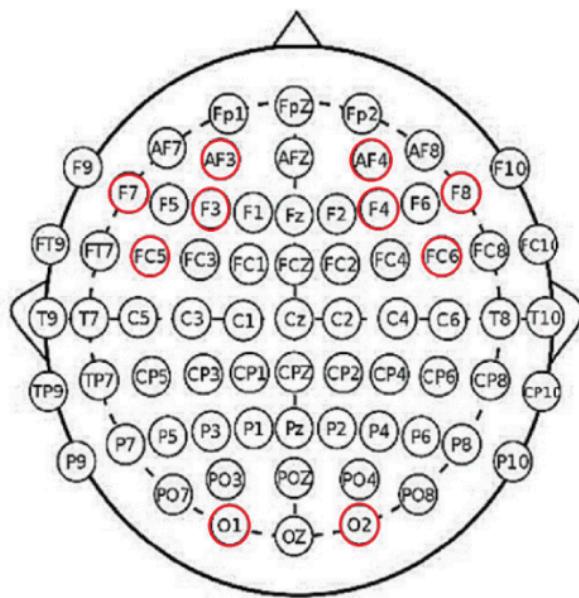


Figura 1: Canales utilizados para medición EEG

Para continuar, se tiene que aplicar la transformada rápida de Fourier para separar los datos en distintas bandas frecuenciales (alfa, beta, gamma, delta, theta) y así poder obtener las potencias relativas. Originalmente, la transformada de Fourier se aplicó para todos los 5 minutos de cada videojuego pero esto resultó en una baja cantidad de datos para la entrada de la red neuronal. Por este motivo, a los 5 minutos de juego se los separó en tramos de 10 segundos a los cuales se les aplicó FFT para así obtener 30 veces más datos.

Una vez obtenidos los datos de las potencias relativas, se aplicó una reducción de dimensionalidad de componentes principales (PCA) sobre las bandas frecuenciales para así obtener una matriz de los canales por la cantidad de componentes principales. Estos son los datos que se utilizaron como entrada a la red neuronal. Se realizó un estudio para determinar qué cantidad de componentes principales es la mejor para alimentar a la red neuronal (de 1 a 5).

2. Modelo

2.1 Layer-wise Relevance Propagation (LRP)

Así como se mencionó anteriormente, las redes neuronales en general se toman como un *black-box*. Se le da una entrada y se obtiene una salida. La manera que la red “elige” la salida es irrelevante. Dicho eso, existen casos en donde entender cómo es que eligió la red lo que eligió, es importante. El campo de *explainable-AI* tiene cada vez mayor relevancia y cuenta con muchos métodos distintos para poder entender las redes neuronales. Para este trabajo práctico, nos enfocaremos en Layer-wise Relevance Propagation (LRP), así como fue definido en *Layer-Wise Relevance Propagation: An Overview*¹.

El método, como su nombre implica, intenta determinar la relevancia de cada neurona en la red a la hora de generar una salida específica. En otras palabras, si una neurona tiene poca relevancia, un cambio de sus pesos sinápticos, tendrá un menor efecto sobre la salida, en comparación a una neurona con mucha relevancia. Cabe nuevamente recalcar que la relevancia es para *una* entrada y salida específica. Para una entrada e_1 , la relevancia de una cierta neurona n_j puede ser muy baja, mientras que para otra entrada e_2 , puede resultar muy alta la relevancia.

El proceso de LRP, se calcula de atrás para adelante, es decir, comienza con la salida de la red y eventualmente llega a la entrada de dicha red. Esta propagación se hace de “capa a capa”, y tiene la particularidad de que la suma de la relevancia de todas las neuronas de una capa da 1. A continuación se enuncia la fórmula para calcular la relevancia de una neurona i en la capa L (existen múltiples versiones, esta es la Z-Rule):

$$R_i = \sum_j \left(\frac{x_i^L w_{ij}^{L+1}}{\sum_k x_k^L w_{kj}^{L+1}} R_j^{L+1} \right) \text{ donde:}$$

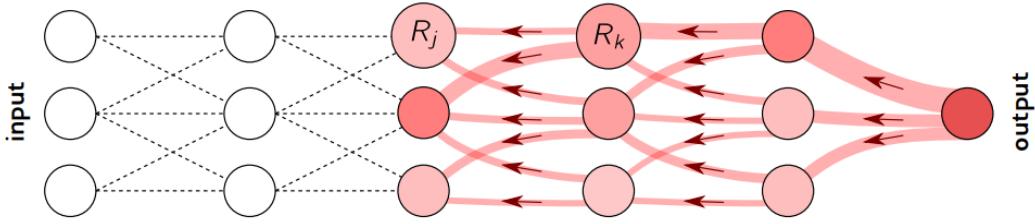
x_i^L : es la salida de la neurona i de la capa L .

w_{ij}^{L+1} : es el peso sináptico entre la neurona i de la capa L y j de la capa $L+1$.

R_j^{L+1} : es la relevancia de la neurona j de la capa $L+1$.

En definitiva, la relevancia de cada neurona se calcula en base a su salida, las salidas de las neuronas de su misma capa, los pesos sinápticos con la capa siguiente y la relevancia de las neuronas de la capa siguiente. Podemos observar esto de forma gráfica:

¹ Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. (n.d.). Layer-Wise Relevance Propagation: An Overview. *Lecture Notes in Computer Science*, 193–209. https://doi.org/10.1007/978-3-030-28954-6_10



²

Figura 2: representación gráfica de LRP para neurona R_j .

Pasando a la implementación que se hizo para este trabajo práctico, se vio necesario hacerlo de forma vectorial, dado que la forma genérica es muy lenta su cálculo. La implementación numérica se define como:

$$R^L = X^L \cdot (W^{L+1} \odot \frac{R^{L+1}}{W^{L+1} X^L + \epsilon})$$

donde R, X, W son los vectores (representativos de una capa entera) de las mismas variables definidas anteriormente, y ϵ un valor pequeño para garantizar la estabilidad numérica.

2.2 Relevancia de conexiones sinápticas

Un aspecto que nos interesa tratar, especialmente cuando empezamos a graficar las neuronas y conexiones, es la relevancia de las *conexiones sinápticas* entre neuronas. Este, no se define en el paper original, por lo que tratamos definir la relevancia de la conexión, en base a la relevancia de las neuronas que conecta. Se definieron 2 reglas distintas:

1. Relevancia promedio: la relevancia conexión se define como el promedio de la relevancia de las 2 neuronas que conecta:

$$P_{ij}^L = \frac{R_i^L + R_j^{L+1}}{2} \quad \text{tal que existe conexión entre } n_i^L \text{ y } n_j^{L+1}$$

2. Relevancia minimizada: usando la relevancia promedio, se le aplica una función la cual solo destaca las mayores relevancias:

$$M_{ij}^L = e^{\alpha \cdot (P_{ij}^L - 1)} - e^{-\alpha} \quad \text{con } \alpha \text{ un factor de ajuste}$$

3. Relevancia decadente: nuevamente se usa la relevancia promedio y se agrega un factor que disminuye su relevancia a medida que se acerca a la capa inicial

$$D_{ij}^L = P_{ij}^L \cdot \frac{L+1}{n_{layers}}$$

² Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. (n.d.). Layer-Wise Relevance Propagation: An Overview. *Lecture Notes in Computer Science*, 193–209. https://doi.org/10.1007/978-3-030-28954-6_10

3. Resultados

Comenzando por la red neuronal, mediante la prueba y error, se encontró que la siguiente configuración daba los mejores resultados (en términos de *accuracy*):

Capa	Cantidad de neuronas	Funcion de activacion
Entrada	$14 \cdot n_{componentes}$	—
Interna 1	64	Relu
Interna 2	32	Relu
Interna 3	16	Relu
Interna 4	8	Relu
Salida	4	Softmax

3.1 Análisis de PCA

Como se mencionó en la introducción, a las potencias relativas se les aplica un análisis de componentes principales sobre las bandas frecuenciales. A partir de esto se tuvo que determinar qué cantidad de componentes principales fue la que mejores resultados obtuvo en términos de *accuracy*. Los valores posibles para este análisis son entre 1, la primera componente principal, y 5, que sería equivalente a usar todos las bandas frecuenciales con una transformación lineal aplicada.

Para cada valor de cantidad de componentes principales se realizó una división de los datos de entrenamiento y de testeo de 80%, realizándose un promedio de 20 ciclos de entrenamiento.

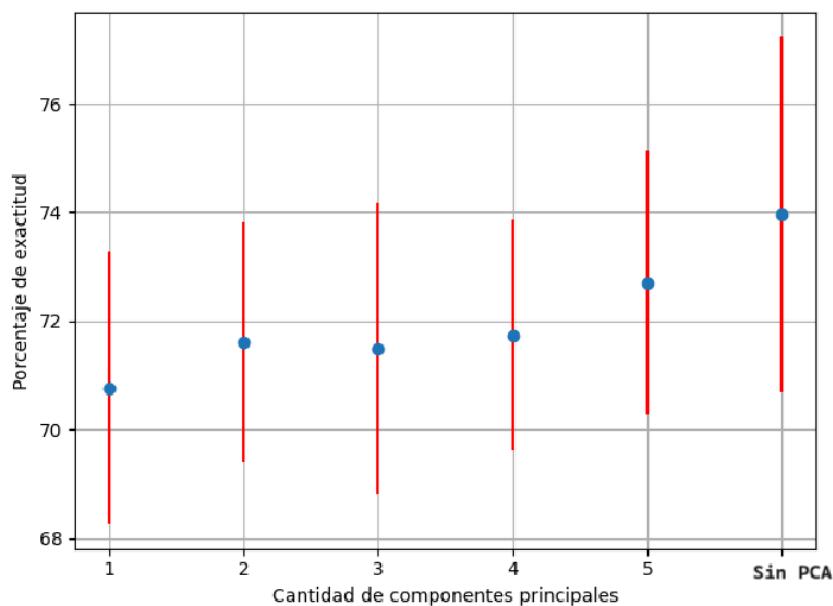


Figura 3: Porcentaje de *accuracy* para las distintas cantidades de componentes principales

En la figura 3 se puede observar que a medida que aumenta la cantidad de componentes principales, el porcentaje de *accuracy* promedio aumenta. Esto se puede atribuir a que, a pesar que se le está agregando información menos relevante (por definición de PCA), sigue siendo más información que la red neuronal puede utilizar. Con 5 componentes principales se obtuvo el porcentaje promedio más alto de 72.7% con el mayor porcentaje de testeo de un solo ciclo de 79.02%.

Un resultado particular fue que con 2 componentes principales se obtuvieron mejores resultados que con 3, más allá de ese caso, parece seguir una relación lineal a medida que aumenta las componentes principales.

De todas formas, las diferencias que presentan los distintos valores de porcentaje no son muy grandes (la diferencia mayor es del 2%) y la desviación estándar es lo suficientemente grande para todos los casos que se encapsulan la totalidad de los otros puntos. De esto se puede decir que no se presentan diferencias significativas.

Por último, algo interesante que sucede es que sin aplicar PCA, se obtienen mejores resultados que cualquier versión de PCA. Uno podría suponer que los datos sin PCA o con la totalidad de los componentes de PCA serían equivalentes ya que la cantidad de datos para la red es la misma pero resulta que la aplicación de la transformación lineal de los datos perjudica al *accuracy* del testeo.

3.2 Análisis de LRP

Para poder analizar la relevancia de la neurona y los pesos sinápticos dividimos los valores de entrada a la red por su valor de salida esperada, es decir, dividimos los datos en función de la clase (aburrido, tranquilo, de horror o divertido) a la cual pertenecían. Teniendo esto, se hacía el análisis LRP usando valores de testeo (es decir, no usadas para el entrenamiento) de entrada aleatorios *de una misma clase*. De esta forma, podríamos observar si, para un valor esperado la relevancia de las neuronas y conexiones se asemejaba para valores de entrada distinta.

Para comenzar, graficamos las neuronas usando una gradiente de color para denotar su relevancia. De la misma forma, los gráficos también incluyen las conexiones sinápticas y la relevancia definida en la sección 2.2. A continuación, se muestran 4 corridas distintas para la clase *aburrido*, habiendo aplicado solo la transformación lineal y usando el método de *relevancia minimizada*:

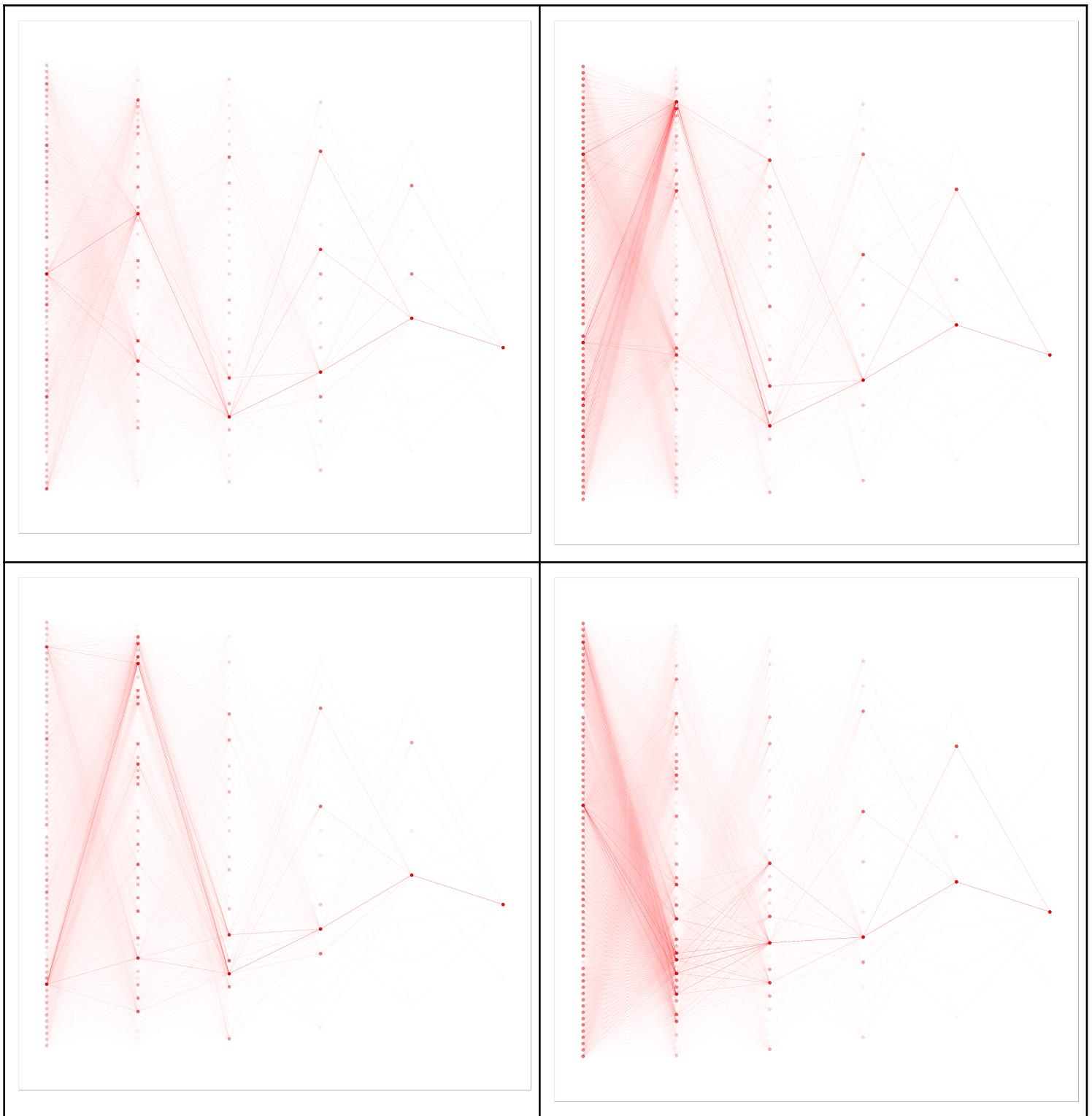


Figura 4: LRP de 4 valores de entrada distintas para la clase *aburrido*
 $(\alpha = 6$ usando método de *relevancia minimizada*)

Como bien se puede observar, existen neuronas que tienen mayor relevancia que otras. Un aspecto sorprendente es que algunas neuronas tienen una relevancia casi nula. Dicho esto, las neuronas de mayor relevancia tienden a repetirse (misma neurona, mismo nivel de relevancia) entre distintas corridas. Parecería indicar que hay una correlación entre la clase elegida por la red y la relevancia *de ciertas neuronas* en dicha red. Este comportamiento

se puede observar en distintas capas de la red, pero es especialmente evidente en las últimas 4 capas de la red. De todas formas, se presentan diferencias significativas en las relevancias en las primeras 2 capas, las cuales parecen distribuir la relevancia más equitativamente entre las neuronas de dicha capa (recordar que la relevancia es por capa y la suma de todas las relevancias de una capa da 1).

Otro aspecto digno de mención, es que si uno las neuronas como nodos y las conexiones sinápticas como aristas, en todos los casos, existe por lo menos 1 camino que vaya de una neurona de entrada a una de salida con una relevancia elevada. Obviamente, en la red las capas son conexas, pero no es obvia que si uno pondera estas conexiones por su relevancia dada por el método LRP, que estas sean conexas o por lo menos, exista una camino distintivo.

Si ahora hacemos un promedio de la relevancia para 20 valores de entrada distintos (incluyendo los 4 usados previamente), podemos graficar esta:

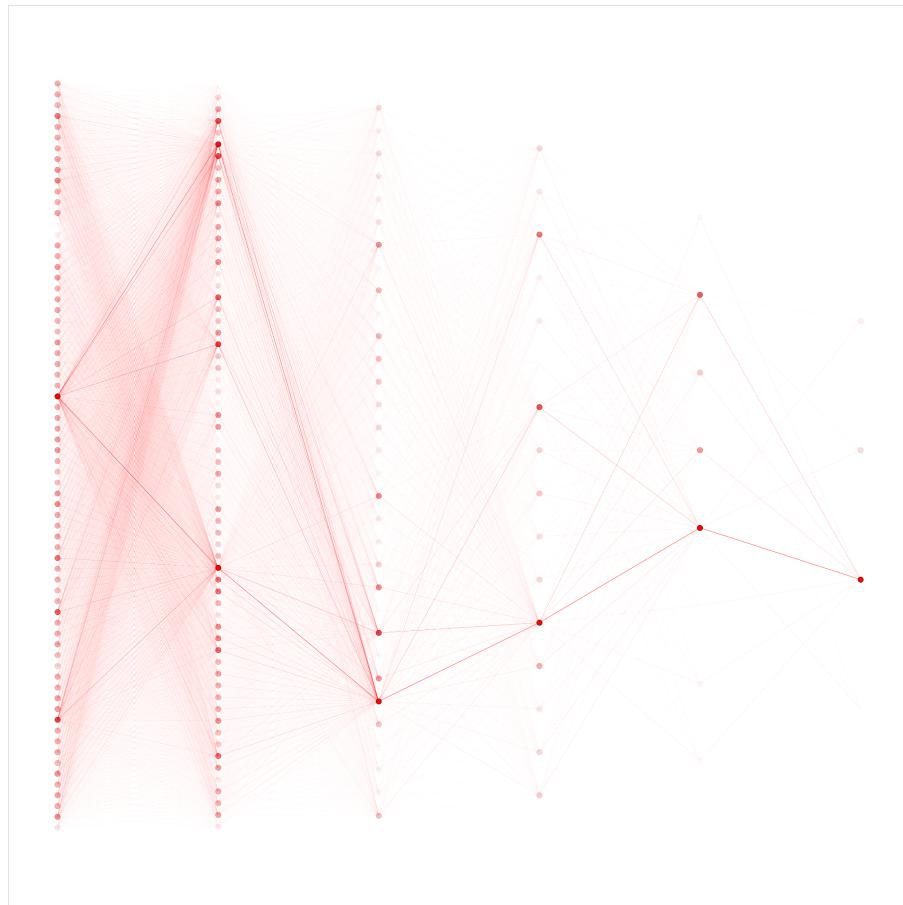


Figura 5: LRP promedio de 20 valores para la clase *aburrido*
($\alpha = 6$ usando método de *relevancia minimizada*)

Si uno contrasta los gráficos individuales de la Fig. 4 con el promedio en la Fig. 5, puede observar que especialmente en la última, penúltima y antepenúltima capa, los valores de relevancia de las neuronas son muy cercanas al promedio, lo cual da más evidencia que las neuronas tienden a tener la misma relevancia en una decisión final, si es que todas corresponden a la elección de una misma clase.

De la misma manera que analizamos el promedio 20 valores de entrada distintos, graficamos también la varianza de la relevancia de las neuronas y sus conexiones sinápticas, observamos un patrón distintivo:

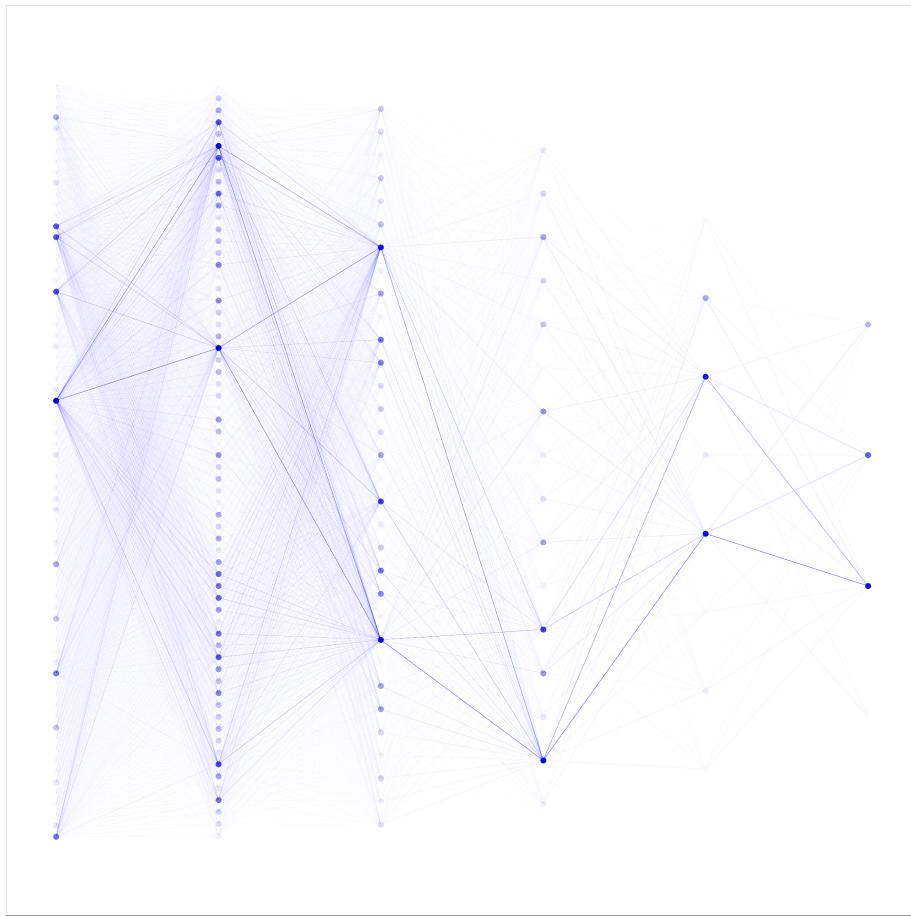


Figura 6: varianza de LRP de 20 valores para la clase *aburrido*
($\alpha = 6$ usando metodo de *relevancia minimizada*)

Podemos observar que las neuronas más relevantes tienden a tener una varianza mayor en muchos de los casos, pero no en todos. De todas formas, la observación más interesante que se obtiene es que, en la primera capa, la mayoría de las neuronas presentan una varianza baja. Hay que tener en cuenta que hay 70 neuronas de entrada y que solo 4 neuronas presentan una varianza alta y otras 8 presentan una varianza mediana, mientras que las restantes 58 neuronas tienen una varianza baja.

La baja variación de las neuronas de entrada da indicios que de, para una “elección” hecha por la red, se tiende a “mirar” las mismas neuronas de entrada. La red efectivamente aprende a distinguir la entrada con el fin de generar una clasificación correcta.

4. Conclusión

Este objetivo de este trabajo fue explorar el concepto de *explainable-AI* e intentar visualizar cómo es que el entrenamiento de la red neuronal afecta a su toma de decisiones para entradas posteriores.

Para explorar esto primero tuvimos que seleccionar un set de datos arbitrarios para poder ver los resultados. Sobre estos datos EEG, se calcularon las potencias relativas para las distintas bandas frecuenciales, y luego se aplicó PCA para ver si seleccionar un set de datos menor, de mayor relevancia, era más efectivo para la red neuronal. De todas formas, pudimos ver que no es así, que es preferible para la red neuronal tener la mayor cantidad de datos. Además, pudimos ver que no aplicar PCA es incluso más útil que aplicarlo con todas las componentes principales, la transformación lineal que se le aplica perjudica a los resultados de la red.

Luego, habiendo entrenado la red neuronal, pudimos aplicar el método de Layer-Wise Relevance Propagation, el cual para una salida generada por la red da la relevancia de cada neurona respecto al resto de las neuronas en su misma capa. Desarrollamos también 3 métodos para discernir la relevancia de las conexiones sinápticas; está siendo: la *relevancia promedio*, la *relevancia minimizada* y la *relevancia decadente*. Junto a esto, pudimos visualizar la relevancia de las neuronas y sus conexiones, y observamos que:

1. Si una red “elige” cierta clase de salida, por lo general, se repetirá el nivel de relevancia de ciertas neuronas para múltiples valores de entrada distintos. Este fenómeno es más prevalente en las últimas capas de la red.
2. Si uno toma la varianza de múltiples aplicaciones del método LRP para valores de entrada distinta, se observa que las neuronas con mayor varianza coinciden, en general, con las de mayor relevancia. También se observó que las neuronas de entrada presentaron baja relevancia.

Dadas estas 2 observaciones, sospechamos que la red neuronal una vez aprendido los características del dataset, al observar suficientes de estas características en un valor de entrada, la propagación de activación tienden a seguir un mismo camino, pasando por las mismas neuronas (las cuales tienen una mayor relevancia en la elección final). No solo eso, respecto a la varianza, esta baja respecto a la relevancia de las neuronas en la primera capa, creemos que esto muestra que la red aprende observar las mismas particularidades de los datos de entrada, en términos simples, “tiende a mirar lo mismo en la data”.

De todas formas, debemos admitir que nuestro conocimiento reducido nos impide hacer un análisis de mayor profundidad. Un aspecto que se podría mejorar es hacer un análisis mayoritariamente cuantitativo (el cual requeriría un mayor conocimiento matemático que el que disponemos), en vez del análisis cualitativo que se hizo. De todas formas, creemos que las observaciones que se pudieron hacer aumentaron nuestro entendimiento de cómo es que las redes neuronales operan.

5. Anexo

5.1 LRP de conexiones sinápticas: método relevancia promediada

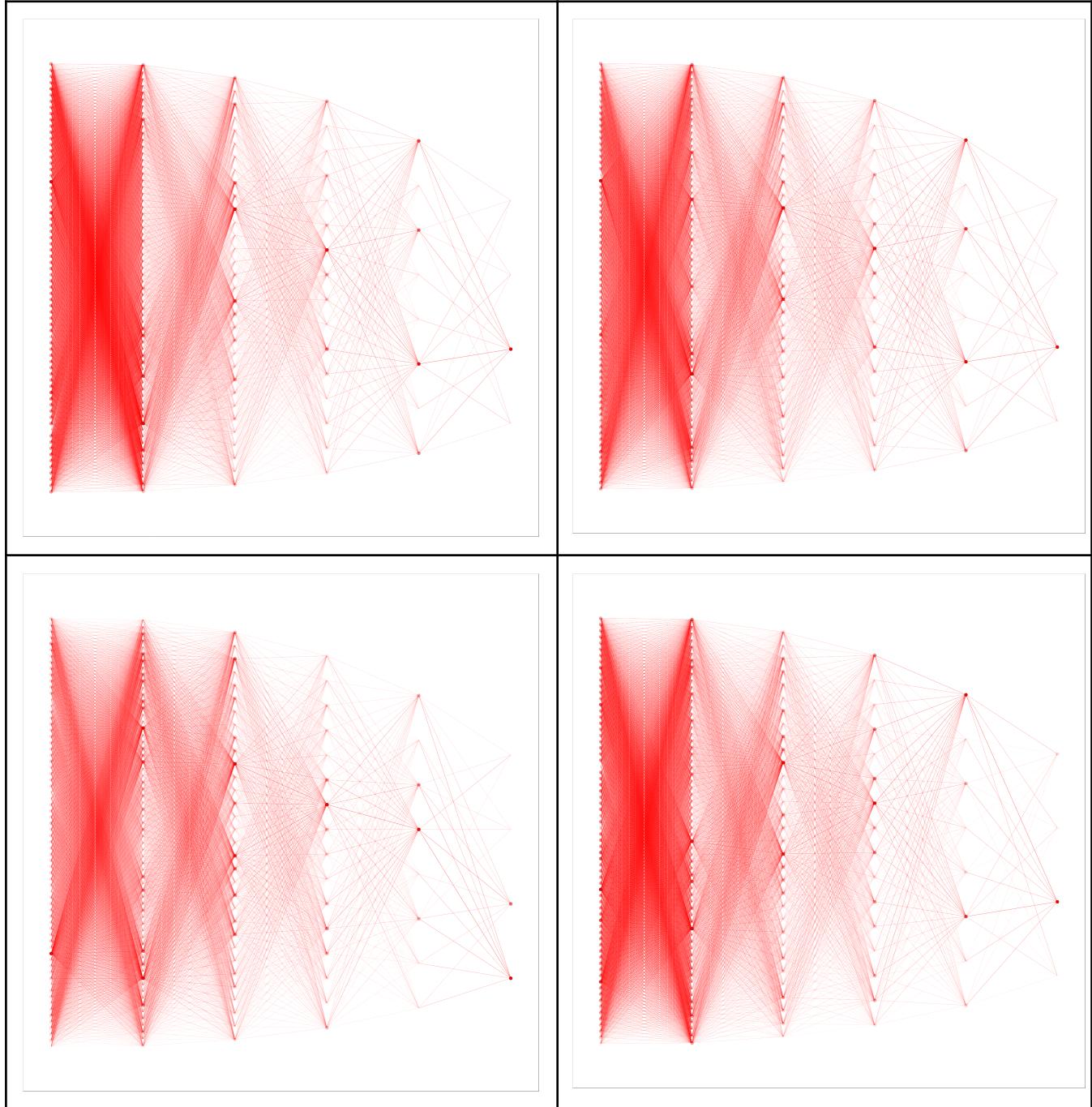


Figura 7: LRP de 4 valores de entrada distintas para la clase *aburrido*
(usando método de *relevancia promediada*)

5.2 LRP de conexiones sinápticas: método relevancia decadente

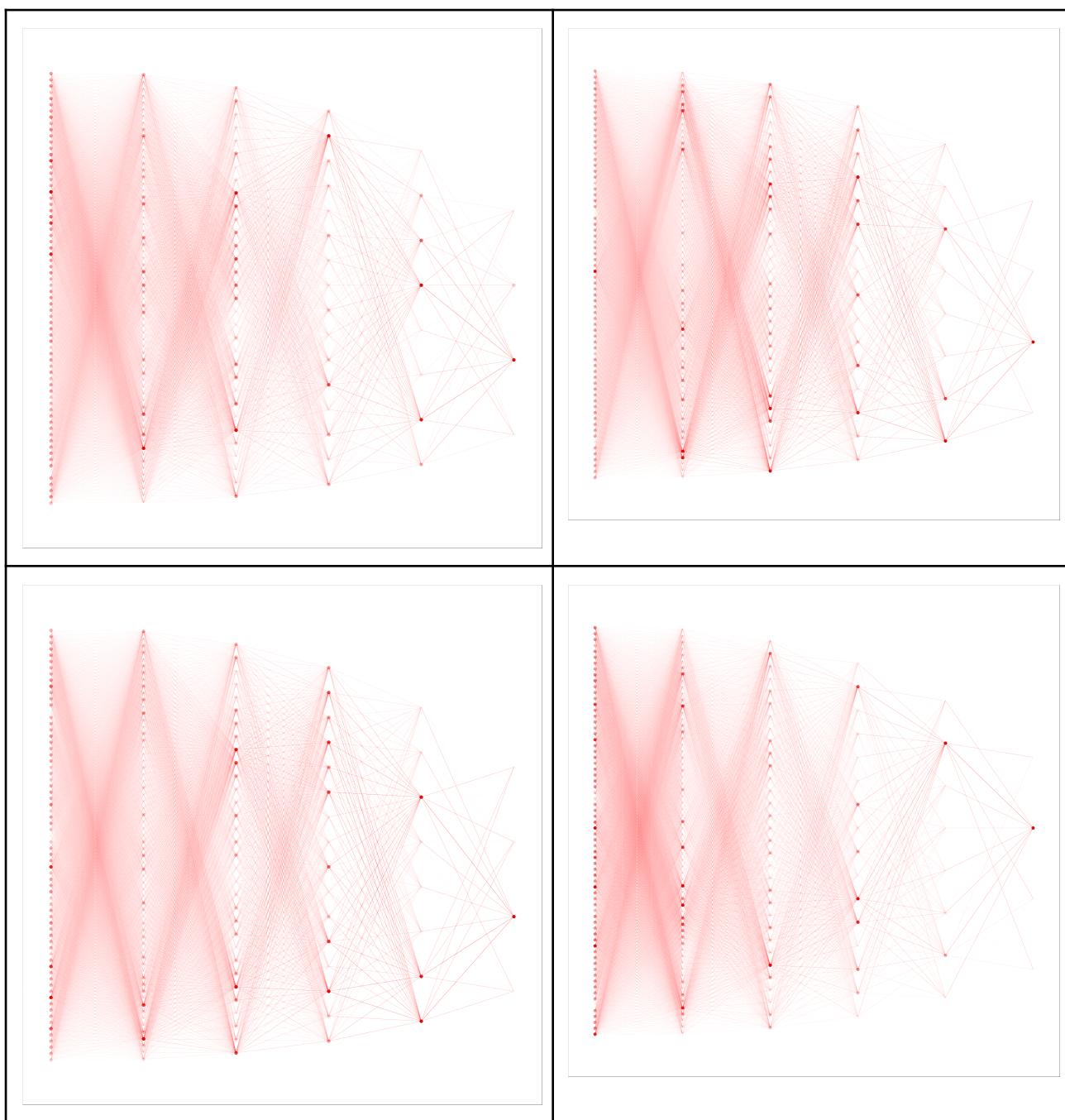


Figura 8: LRP de 4 valores de entrada distintas para la clase *aburrido*
(usando método de *relevancia decadente*)

5.3 LRP de valores de entrada de 4 clases distintas

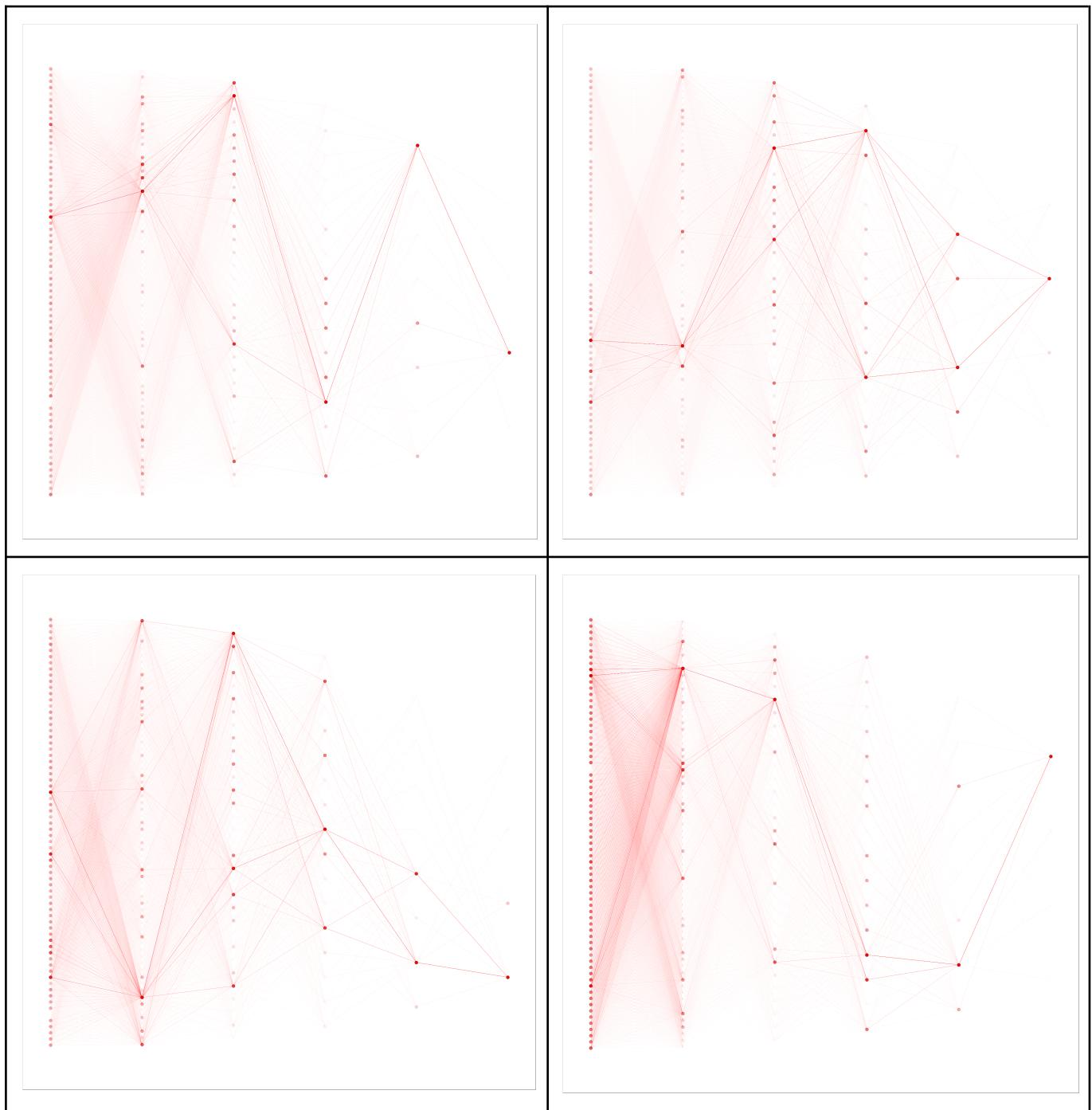


Figura 9: LRP de 4 valores de entrada distintas para la clases *aburrido*, *tranquilo*, *de horror* o *divertido* ($\alpha = 6$ usando método de *relevancia minimizada*)