Título de la revista
Revista **XX** (XXXX) XXXXXX
https://doi.org/XXXX/XXXX

Red neuronal 4D basada en la atención para el reconocimiento de emociones por EEG

Guowen Xiao [1] - Mengwen Ye [2] - Bowen Xu [1] - Zhendi Chen [1] - Quansheng Ren [1*]

Resumen

El reconocimiento de emociones mediante electroencefalograma (EEG) es una tarea importante en el campo de la interfaz cerebro-ordenador. Aunque recientemente se han propuesto muchos métodos de aprendizaje profundo, sigue siendo difícil aprovechar al máximo la información contenida en los diferentes dominios de las señales de EEG. En este artículo, presentamos un método novedoso, denominado red neuronal de cuatro dimensiones basada en la atención (4D-aNN) para el reconocimiento de emociones mediante EEG. En primer lugar, las señales EEG en bruto se transforman en representaciones espacio-espectrales-temporales 4D. A continuación, la 4D-aNN propuesta adopta mecanismos de atención espectral y espacial para asignar de forma adaptativa los pesos de diferentes regiones cerebrales y bandas de frecuencia, y se utiliza una red neuronal convolucional (CNN) para tratar la información espectral y espacial de las representaciones 4D. Además, se integra un mecanismo de atención temporal en una memoria a corto plazo bidireccional (LSTM) para explorar las dependencias temporales de las representaciones 4D. Nuestro modelo alcanza un rendimiento puntero en el conjunto de datos SEED con división intra-sujeto. Los resultados experimentales han demostrado la eficacia de los mecanismos de atención en diferentes dominios para el reconocimiento de emociones en EEG.

Palabras clave: EEG, reconocimiento de emociones, mecanismo de atención, red neuronal convolucional recurrente

[1] Departamento de Electrónica, 1

Universidad de Pekín, Pekín, China

[2] Escuela de Ingeniería Eléctrica,

Universidad Jiaotong de Pekín, Pekín, China

*Autor correspondiente: Quansheng Ren (Correo electrónico: qsren@pku.edu.cn)

Introducción

La emoción desempeña un papel importante en la vida cotidiana y está estrechamente relacionada con el comportamiento humano y la cognición (Dolan 2002). Como uno de los temas de investigación más significativos de la computación afectiva, el reconocimiento de emociones ha recibido una atención creciente en los últimos años por su aplicación s de detección de enfermedades (Bamdad et al. 2015; Figueiredo et al. 2019) , la interacción persona-ordenador (Fiorinia et al. 2020; Katsigiannis y Ram zan 2017), y la estimación de la carga de trabajo (Blankertz et al. 2016). En general, los métodos de reconocimiento de emociones pueden dividirse en dos categorías (Mühl et al. 2014). Una se basa en las respuestas emocionales externas, incluidas las expresiones faciales y los gestos (Yan et al. 2016), y la otra se basa en las respuestas emocionales internas, incluidos el electroencefalograma (EEG) y el electrocardiograma (ECG) (Zheng et al. 2017). Las investigaciones neurocientíficas han demostrado que algunas regiones importantes de la corteza cerebral están estrechamente relacionadas con las emociones, por lo que es posible decodificar las emociones basándose en el EEG (Brittona et al. 2006; Lotfia y Akbarzadeh -T 2014). El EEG no es invasivo, es portátil y barato, por lo que se ha utilizado ampliamente en el campo de las interfaces cerebro-ordenador (BCI) (Pfurtscheller et al. 2010). Además, las señales EEG contienen diversa información espacial, espectral y temporal sobre las emociones evocadas por patrones de estimulación específicos. Por lo tanto, cada vez más investigadores se centran en el reconocimiento de emociones EEG recientemente (Alhagry et al. 2017; L i y Lu 2009).

Los métodos tradicionales de reconocimiento de emociones por EEG suelen extraer primero características artesanales de las señales de EEG y luego adoptan modelos poco profundos para clasificar las características de la emoción. Las características emocionales del EEG pueden extraerse del dominio temporal, del dominio de la frecuencia y del dominio de la frecuencia temporal. Jenke et al. realizan un estudio exhaustivo sobre los métodos de extracción de características del EEG mediante técnicas de aprendizaje automático en un conjunto de datos autograbados (Jenke et al. 2014). Para clasificar las características de emoción extraídas, muchos investigadores han adoptado métodos de aprendizaje automático en los últimos años (Kim et al. 2013). Li et al. aplican una máquina lineal de vectores de soporte (SVM) para clasificar las características emocionales extraídas de la banda de frecuencia gamma (Li y Lu 2009). Duan et al. extraen características de entropía diferencial (DE), que son superiores para representar estados de emoción en señales de EEG (Shi et al. 2013), de datos de EEG multicanal y combinan un k -Nearest Neighbor (KNN) con SVM para clasificar las características DE (Duan et al. 2013). Sin embargo, los modelos superficiales requieren mucho conocimiento experto para diseñar y seleccionar las características de emoción, lo que limita su rendimiento en la clasificación de emociones EEG.

Se ha demostrado que los métodos de aprendizaje profundo superan a los métodos tradicionales de aprendizaje automático en muchos campos como la visión por ordenador, el procesamiento del lenguaje natural y el procesamiento de señales biomédicas (Abbass et al. 2018; Craik et al. 2019) por la capacidad de aprender características de alto nivel a partir de datos automáticamente (Krizhevsky et al. 2012). Recientemente, algunos investigadores han aplicado el aprendizaje profundo a la emoción EEG

reconocimiento. Zheng et al. introducen una red de creencia profunda (DBN) para investigar las bandas de frecuencia críticas y los canales de señal EEG para el reconocimiento de emociones EEG (Zheng y Lu 2015). Yang et al. proponen una red jerárquica para clasificar las características de EEG extraídas de diferentes bandas de frecuencia (Yang et al. 2018b). Song et al. utilizan una red neuronal convolucional gráfica para clasificar las características DE (Song et al. 2020). Ma et al. proponen un modelo de memoria residual multimodal a corto plazo (MMResLSTM) para el reconocimiento de emociones, que comparte pesos temporales entre las múltiples modalidades (Jiaxin Ma et al. 2019). Para aprender la discrepancia bi-hemisférica para el reconocimiento de emociones EEG, Yang et al. proponen un nuevo modelo de discrepancia bi-hemisférica (BiHDM) (Li et al. 2020). Todos estos métodos de aprendizaje profundo superan a los modelos superficiales.

Aunque los modelos de reconocimiento de emociones de aprendizaje profundo han logrado una mayor precisión que los modelos superficiales, sigue siendo un reto fusionar información más importante en diferentes dominios y capturar patrones locales discriminativos en las señales de EEG. En las últimas décadas, muchos investigadores han estudiado las bandas de frecuencia y los canales críticos para el reconocimiento de emociones en EEG. Zheng et al. demuestran que las bandas β[14~31 Hz] y γ[31~51 Hz] están más relacionadas con el reconocimiento de emociones que otras bandas, y su modelo consigue el mejor rendimiento al combinar todas las bandas de frecuencia. También realizan experimentos para seleccionar los canales críticos y proponen los grupos mínimos de conjuntos de electrodos para el reconocimiento de emociones (Zheng y Lu 2015). Para utilizar la información espacial de las señales de EEG, Li et al. proponen un mapa disperso 2D para mantener la información oculta en la colocación de electrodos (Li et al. 2018). Zhong et al. introducen una red neuronal gráfica regularizada (RGNN) para capturar las relaciones locales y globales entre los diferentes canales de EEG para el reconocimiento de emociones (Zhong et al. 2020). Las dependencias temporales en las señales EEG también son importantes para el reconocimiento de emociones. Por ejemplo, Ma et al. (Jiaxin Ma et al. 2019) aplican LSTMs en sus modelos para extraer características temporales para el reconocimiento de emociones. Shen et al. transforman las características DE de diferentes canales en estructuras 4D para integrar la información espectral, espacial y temporal simultáneamente y luego utilizan una red neuronal recurrente convolucional de cuatro dimensiones (4D - CRNN) para reconocer diferentes emociones (Shen et al. 2020). Sin embargo, las diferencias entre regiones cerebrales y bandas de frecuencia no se utilizan plenamente en su trabajo. Para capturar de forma adaptativa patrones discriminativos en las señales EEG, se han aplicado mecanismos de atención al reconocimiento de emociones EEG. Por ejemplo, Tao et al. introducen un mecanismo de atención por canales, asignando los pesos de diferentes canales de forma adaptativa, junto con una autoatención extendida para explorar las dependencias temporales de las señales EEG (Tao et al. 2020). Jia et al. proponen una red de dos flujos con mecanismos de atención para centrarse de forma adaptativa en patrones importantes (Jia et al. 2020). De lo anterior se desprende que es fundamental

para integrar información de diferentes dominios y capturar de forma adaptativa regiones cerebrales importantes, bandas de frecuencia y marcas de tiempo en una red unificada para el reconocimiento de emociones EEG. En este trabajo, proponemos una red neuronal cuatridimensional basada en la atención denominada 4D-aNN para el reconocimiento de emociones EEG. En primer lugar, transformamos las señales EEG en bruto en representaciones espacio-espectrales-temporales 4D que consisten en varios cortes temporales. Las distintas regiones cerebrales y bandas de frecuencia varían en su contribución al reconocimiento de emociones por EEG, por lo que también deben tenerse en cuenta las dependencias temporales de las representaciones 4D. Por lo tanto, empleamos mecanismos de atención tanto en una CNN como en una red LSTM bidireccional para capturar de forma adaptativa patrones discriminativos. Para el modelo CNN, el mecanismo de atención se aplica a las dimensiones espaciales y espectrales de cada corte temporal, de modo que se puedan capturar las regiones cerebrales y las bandas de frecuencia importantes. En cuanto al modelo LSTM bidireccional, el mecanismo de atención se aplica para utilizar las dependencias temporales de largo alcance, de modo que la importancia de los diferentes patrones temporales se refleje en el modelo CNN. rebanadas en una representación 4D podrían explorarse por completo.

Las principales aportaciones de este trabajo se resumen como sigue: a) Proponemos una red neuronal cuatridimensional basada en la atención, que fusiona información de diferentes dominios y captura patrones discriminatorios en señales EEG basadas en la representación espacio-espectral-temporal 4D. b) Llevamos a cabo experimentos en el conjunto de datos SEED, y los resultados experimentales indican que nuestro modelo alcanza un rendimiento de vanguardia en la división intra-sujeto.

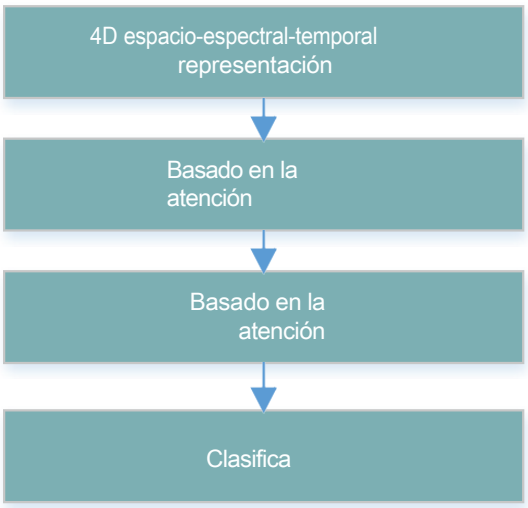
El resto de este documento se organiza como sigue. En la sección *Método* se describe el método propuesto. El conjunto de datos, la configuración del experimento, los resultados, los estudios de ablación y la discusión se presentan en la sección *Experimento*. Por último, en la sección *Conclusiones se* exponen las conclusiones.

Método

La figura 1 ilustra la estructura general de 4D -aNN para el reconocimiento de emociones EEG. Consiste en la representación 4D espacial-espectral-temporal, la CNN basada en la atención, la LSTM bidireccional basada en la atención y el clasificador. En

describirá los detalles de cada parte en

secuencia.



F ig. 1 La estructura general de 4D-aNN.

Representación espacio-espectral-temporal 4D

El proceso de generación de la representación 4D se muestra en la Fig. 2. Al igual que en trabajos anteriores (Shen et al. 2020; Yang et al. 2018a), dividimos las señales EEG originales en segmentos de T segundos de duración sin solapamiento. A cada segmento se le asigna la misma etiqueta que a las señales de EEG originales. A continuación, descomponemos cada segmento en cinco bandas de frecuencia (es decir, δ[1~4 Hz], θ[4~8 Hz], α[8~14 HZ], β[14~31 Hz] y γ[31~51 Hz]) con filtros Butterworth de cinco órdenes. Las características de entropía diferencial (DE) y densidad espectral de potencia (PSD) de todos los canales de EEG, que han demostrado ser eficaces para el reconocimiento de emociones (Zheng et al. 2017), se extraen de cinco bandas de frecuencia respectivamente con una ventana de 0,5 s para cada segmento.

La PSD se define como

$$_{P}h\,(X) = E[x^2\,] \qquad (1)$$

donde x es formalmente una variable aleatoria y en este contexto , la señal adquirida de una determinada banda de frecuencia en un determinado canal de EEG.

La característica DE es capaz de discriminar patrones de EEG entre energía de baja y alta frecuencia, que se define como

$$hD(X) = -\int_{\square}^{\infty} f(x)\,\log(f(x))\,dx \qquad (2)$$

donde □(□) es la función de densidad de probabilidad de □ . Si □ obedece a la distribución gaussiana □(□, □²), DE puede calcularse simplemente mediante la siguiente formulación:

$$_{D}h(X) = -\int \frac{1}{\sqrt{2\pi\sigma^2}} \frac{(x-\mu)^2}{2\sigma^2} \log \frac{1}{\sqrt{2\pi\sigma^2}} \frac{(x-\mu)^2}{2\sigma^2} \quad (3)$$

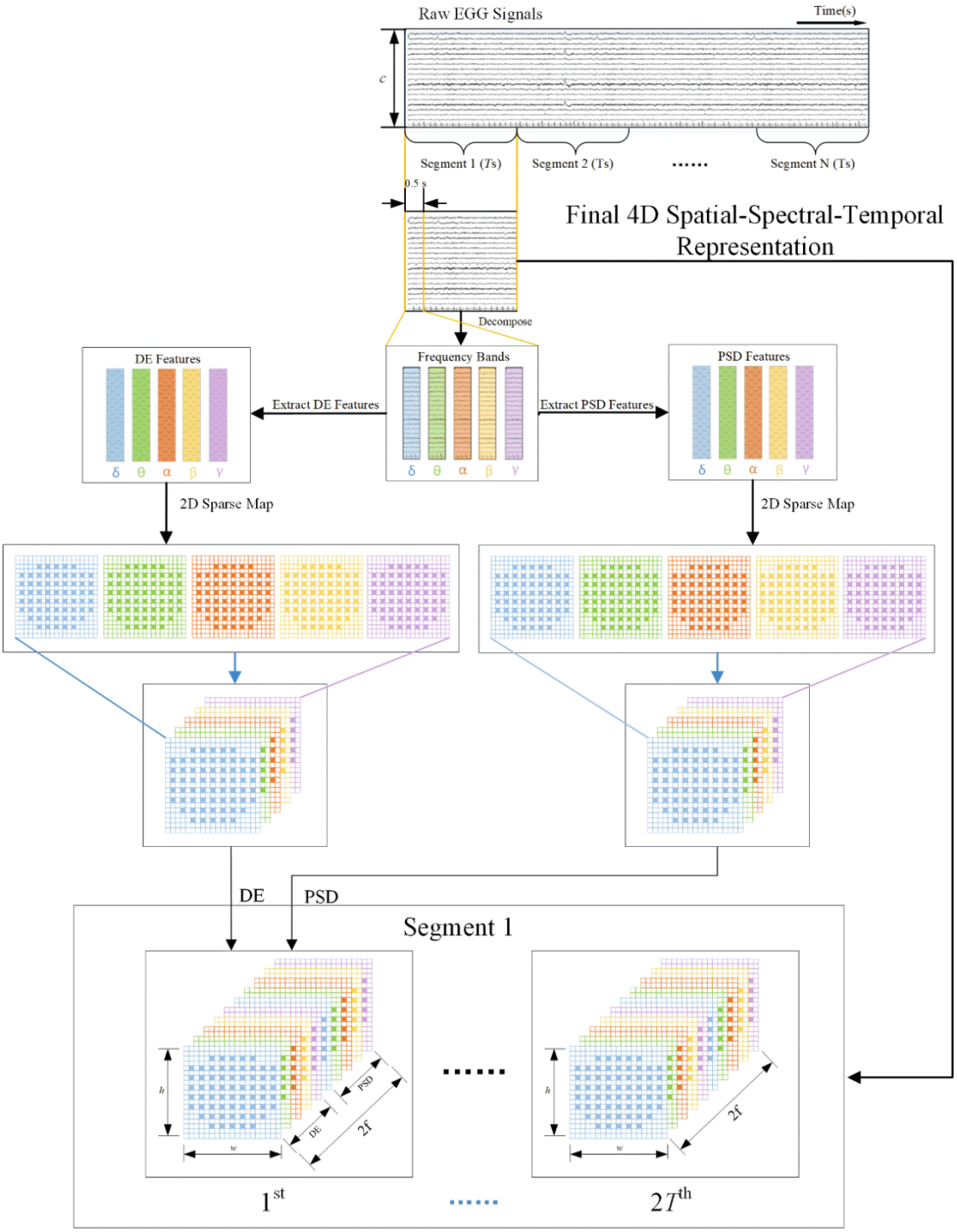
$$= \frac{1}{2}\log 2\pi e\sigma^2$$

donde e y σ son la constante de Euler y la desviación estándar de □, respectivamente.

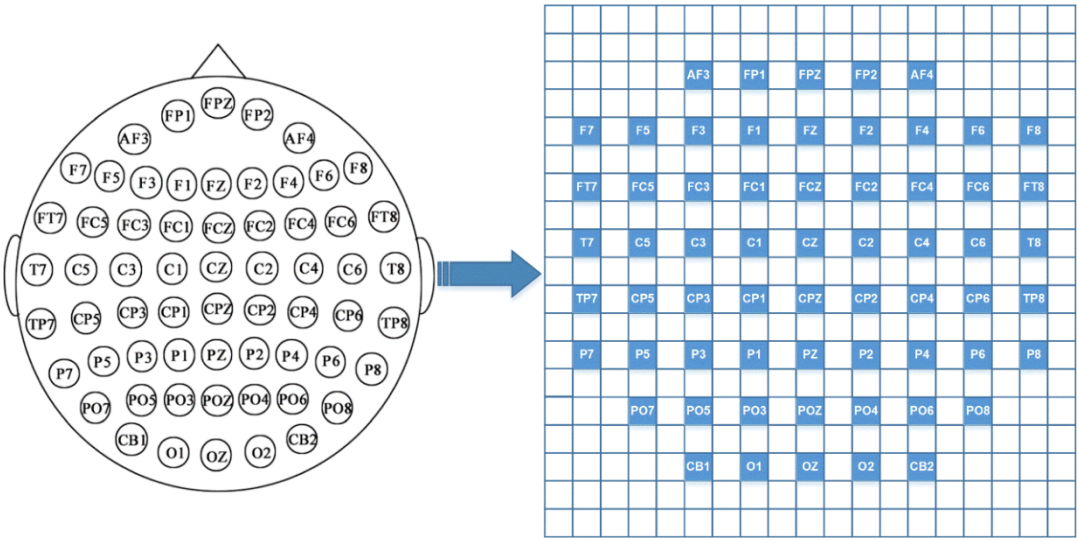
Así, extraemos un tensor de características 3D $F_n \in R^{c\times 2f\times 2T}$, $n = 1, 2, \ldots, N$ de cada segmento, donde N es el número de segmentos totales, c es el número de canales EEG, $2f$ representa las características DE y PSD de f bandas de frecuencia, y $2T$ es

derivados por la ventana de 0,5s sin solapamiento. Para utilizar la información espacial de los electrodos, organizamos todos los canales c como un mapa 2D disperso de modo que el tensor de características 3D F_n se transforme en una representación 4D $X_n \in R^{h \times w \times 2f \times 2T}$, donde h y w son la altura y la anchura del mapa 2D disperso.

respectivamente. El mapa 2D disperso de todos los canales c con relleno cero se muestra en la Fig. 3, que conserva la topología de los diferentes electrodos. En este trabajo, fijamos h = 19, □ = 19, y □ = 5.



F ig. 2 Generación de la representación espacio-temporal 4D. Para cada segmento de señal EEG Ts, extraemos características DE y PSD de diferentes canales s y bandas de frecuencia con una ventana de 0,5s. A continuación, las características se transforman en una representación 4D que consta de cortes temporales de *2T*.

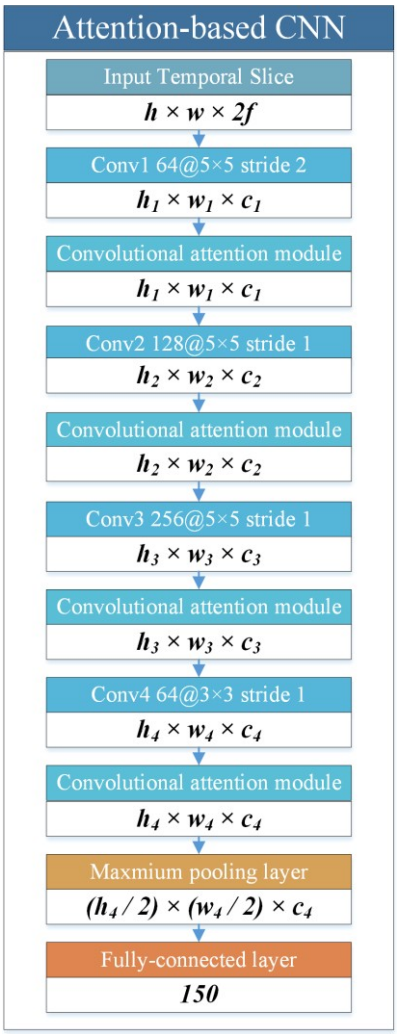


F ig. 3 El mapa 2D disperso con relleno cero de 62 canales. El propósito de la organización es preservar las relaciones posicionales entre los diferentes electrodos.

CN basada en la atención N

Para una representación espacio-espectral-temporal 4D X_n , extraemos la información espacial y espectral de cada corte temporal $S_i \in R^{h \times w \times 2f}$, i = 1, 2, . . . , $2T$ con una CNN, explorar los patrones locales discriminativos en los dominios espacial y espectral con un módulo de atención convolucional , y finalmente obtener su representación espacial y espectral . El módulo de atención aquí es similar a lo que proponen Woo et al. (Woo et al. 2018), que se utiliza originalmente para mejorar el poder de representación de las redes CNN.

La estructura de la CNN basada en la atención se muestra en la Fig. 4. Contiene cuatro capas convolucionales, cuatro módulos de atención convolucionales, una capa de agrupamiento máximo y una capa totalmente conectada. Las cuatro capas convolucionales tienen 64, 128, 256 y 64 mapas de características con un tamaño de filtro de 5×5, 5×5,
5×5 y 3×3, respectivamente. En concreto, se utiliza un módulo de atención convolucional después de cada capa convolucional para utilizar el mecanismo de atención espacial y espectral s, y los detalles se darán más adelante . Sólo utilizamos una capa de max-pooling con un tamaño de filtro de 2×2 después del último módulo de atención convolucional para preservar más información y mejorar la robustez de la red. Por último, las salidas de la capa de agrupamiento máximo se aplanan y se alimentan a la capa totalmente conectada con 150 unidades. Así, para cada sli ce temporal S_i , tomamos la salida final $P_i \in R^{150}$ como su representación espacial y espectral.



F ig. 4 Estructura de la CNN basada en la atención. La mitad

superior de los bloques de la figura es el tipo de capas y la inferior denota la forma de sus tensores de salida.

Módulo de atención convolucional

El módulo de atención convolucional se aplica después de cada capa convolucional para capturar de forma adaptativa regiones cerebrales y bandas de frecuencia importantes. En la Fig. 5 se muestra la estructura del módulo de atención convolucional. Consta de dos submódulos: el módulo de atención espacial y el módulo de atención espectral.

Para cada capa convolucional anterior, su resultado es un tensor de características 3D $V \in R^{h_v \times w_v \times \square\square}$, donde $h\square$, w_v , y c_v son la altura de los mapas de características 2D de $\square$, la anchura de los mapas de características 2D de $\square$, y el número de los mapas de características 2D de $\square$, respectivamente. Tomamos V como entrada del módulo de atención convolucional.

El módulo de atención espectral se aplica para identificar bandas de frecuencia valiosas para el reconocimiento de emociones. La agrupación de medias se ha utilizado ampliamente para agregar información espacial y la
La agrupación máxima se ha adoptado comúnmente para reunir
rasgo distintivo s. Por lo tanto, reducimos el espacio
dimensión de $\square$ mediante una agrupación espacial media y una
máxima agrupación espacial, que se definen como:

$$\square_{\square\square} = \frac{1}{h \times w_v} \sum_{h=1}^{h_v} \sum_{w=1}^{w_v} V_i(h, w) \text{ , } i = 1, 2, \ldots, c \quad (4)$$

$$\square\square\square\square,\square = \square\square\square(\square_{\square}), \square = 1, 2, \ldots, \square_{\square} \quad (5)$$

donde $V_i \in R^{h_v \times w_v}$ denota el mapa de características 2D en la *i-ésima*
canal de $\square$, $C_{avg,i}$ representa el elemento en el *i-ésimo* canal de la representación media espacial $C_{avg} \in R^{c_v}$, $max(Z)$ devuelve el elemento mayor en $\square$, y $C_{max,i}$ es el elemento en el *i-ésimo* canal de la representación espacial máxima. $\square\square\square\square \in R^{c}$ $\square$. Posteriormente, implementamos la atención espectral mediante dos capas totalmente conectadas, una función de activación *Relu* y una función de activación *sigmoidea*, que se define como:

$$\square\square\square\square\square\square\square\square\square,\square\ \square\ \square_2 = \square\square(\square_1\square\square_{\square\square}(\square\square\square\square\square\square\square)) \quad (6)$$

$$\square\square\square\square\square\square\square\square\square\square,\square_{\square\square\square} = \square_{\square\square}(\square\square_{\square\square}(\square\square\square\square\square\square\square)) \quad (7)$$

correspondientes

$$A_{spectral} = sigmoid(A_{spectral,avg} \oplus A_{spectral,max}) \quad (8)$$

$$\square\square\square\square(\square) = \max(\square, 0) \quad (9)$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (10)$$

donde W^S y W^S son parámetros aprendibles, $\oplus$ denota la suma de elementos, y $A_{spectral} \in R^{1 \times 1 \times c_v}$ es la atención espectral. Los elementos de $A_{spectral}$ representan la importancia de los correspondientes mapas de características 2D del dominio espectral. Después de generar la atención espectral $A_{spectral}$, la salida del módulo de atención espectral puede definirse como:

$$V' = A_{spectral} \otimes V \quad (11)$$

donde V denota el tensor de características 3D refinado, y $\otimes$
representa la multiplicación por elementos.

El módulo de atención espacial se aplica para identificar regiones cerebrales valiosas para el reconocimiento de emociones. En primer lugar, reducimos la dimensión espectral de V mediante la agrupación de medias espectrales.
y la agrupación máxima espectral, que se define como:

$$SPAa_{vg,(h,w)} = \frac{1}{\square} \sum_{\square=1}^{\square\square} S'_{h,w}(c) \text{ , } h = 1, 2, \ldots, h_v \text{ ; } w = 1, 2, \ldots, \square_{\square} \quad (12)$$

$$SPAmax_{,(h,w)} = max(S'_{h,w}), h = 1, 2, \ldots, h_v \text{ ; } w = 1, 2, \ldots, \square_{\square} \quad (13)$$

donde $S'_{h,w}$ denota el canal en la fila *h-ésima* y *w-ésima*
columna de V', $SPA_{avg,(h,w)}$ representa el elemento en el *h-ésimo*
fila y columna *w-ésima* de la representación espectral media $SPAavg \in R^{h_v \times w_v \times 1}$ y $SPAmax_{,(h,w)}$ es el elemento en el *h-*
fila *w* y columna *w* del máximo espectral representación $SPAmax \in R^{h_v \times w_v \times 1}$. En lo que sigue

implementan la atención espacial con una capa convolucional y una función de activación *sigmoidea*, que se define como:

$$\square\square\square = \square\square\square(\square\square\square_{\square\square\square}, \square\square\square\square\square) \quad (14)$$

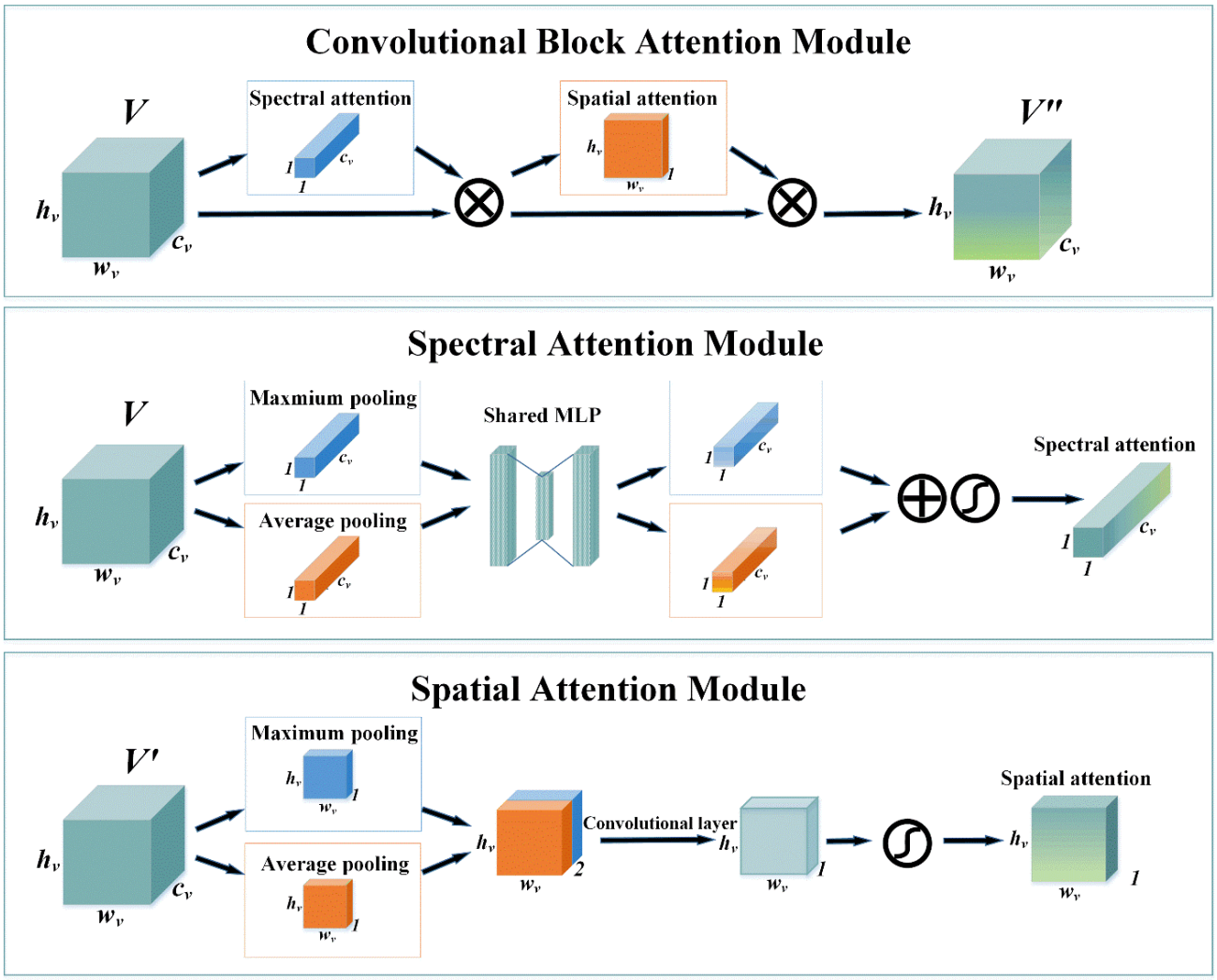
$$\square\square\square\square\square\square\square\square = \square\square\square\square\square\square\square(\square\square\square\square(\square\square\square)) \quad (15)$$

donde $\square\square\square(\square\square\square_{\square\square\square}, \square\square\square\square\square)$ denota la concatenación de
$\square\square\square\square_{\square\square}$ y $\square\square\square\square\square$ a lo largo de la espectral,
$\square\square\square\square(\square\square\square)$ representa la capa convolucional para $\square\square\square$, y
$\square_{\square\square} \in R^{h_v \times w_v \times 1}$ es la atención espacial. Los elementos de $\square\square\square\square\square\square\square\square$ representan la importancia de las regiones

del dominio espacial. Posteriormente, la salida del dominio espacial
módulo de atención puede definirse como:

$$V'' = A_{spatial} \otimes V \quad (16)$$

donde $V'' \in R^{h_v \times w_v \times c_v}$ denota el tensor de características 3D de salida final del módulo de atención convolucional.



F ig. 5 El bloque superior es la estructura general del bloque de atención convolucional, está formado por el módulo de atención espectral y el módulo de atención espacial. El bloque central representa la generación de la atención espectral. El bloque inferior representa la generación de la atención espacial.

STM bidireccional basado en la atención

Para cada tramo temporal $S_i \in R^{h \times w \times 2f}$, i = 1, 2, . . . , 2□, el

El resultado final de la CNN basada en la atención es $P \in R^{150}$. Dado que la

La variación entre diferentes cortes temporales contiene información temporal para el reconocimiento de emociones, utilizamos un LSTM bidireccional basado en la atención para explorar la importancia de los diferentes cortes, como se muestra en la Fig. 6.

Una LSTM bidireccional conecta dos LSTM unidireccionales con direcciones opuestas a la misma salida. En comparación con una LSTM unidireccional, una LSTM bidireccional conserva información tanto del pasado como del futuro, lo que le permite comprender mejor el contexto. En este trabajo, la LSTM bidireccional se compone de dos LSTM unidireccionales con 36 celdas de memoria. La LSTM unidireccional para dirección temporal positiva , LSTM toma la secuencia de salida de la CNN basada en atención □□ = (□ , □ , . . ., □) como secuencia de entrada, mientras que la otra para

^{N}P= (P_{2T} , P_{2T-1} , . . ., P_1) como secuencia de entrada. Las salidas del nodo *i-ésimo* de las LSTM unidireccionales son $Y^P \in R^{36}$ y $Y^N \in R^{36}$,i= 1, 2, . . . , 2T , respectivamente. A continuación, concatenamos Y^P y Y^N como salida del nodo *i-ésimo*

del LSTM bidireccional $Y_i \in R^{72}$. A diferencia de las formas tradicionales que sólo utilizan la salida del último nodo de un LSTM para la clasificación u otras aplicaciones , tomamos las salidas de todos los nodos LSTM bidireccionales $Y \in R^{2T \times 72}$ en consideración y exploramos la importancia de las diferentes rebanadas temporales por el mecanismo de atención temporal.

El mecanismo de atención temporal se implementa con dos capas totalmente conectadas, una función de activación *Relu* y una función de activación *softmax*, que se define como:

□□□□ = □□ (□□□□(□ Y□ + □□))) + □□ (17)

□□□□□□□□ □= □ □ □ □ □ □ □(□□□) (18)

□□
□(□
)

$$softmax(x) = \frac{}{\sum exp(x)} \quad (19)$$

dirección temporal negativa, LSTMN toma la secuencia inversa

donde □1□ , □2□ , □1□ , y □2□ son parámetros aprendibles, □□□□ función para predecir la etiqueta de la muestra 4D □□, que puede

representa el elemento *i-ésimo* de $Tem \in R^{2T\times 1}$ que proyecta

□ ∈ $R^{2T\times 72}$ a una dimensión inferior, y $A_{temporal} \in R^{2T\times 1}$ es la atención temporal. Los elementos de □□□□□□□□□□ representan

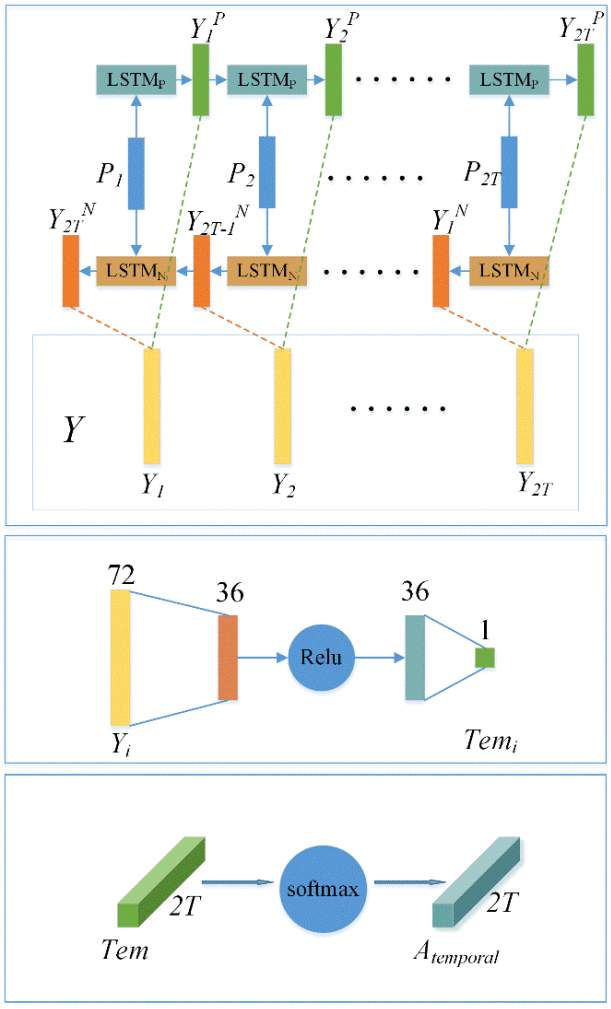
la importancia de los cortes temporales correspondientes. Posteriormente, la representación de alto nivel de la muestra 4D

□ □puede definirse como:

$$_{n}L(e) = \sum A_{temporal} \otimes Y_e \ , \ e = 1, 2, \dots , 72 \quad (20)$$

donde $Y_e \in R^{2T\times 1}$ denota la columna *e-ésima* de $Y \in R^{2T\times 72}$ y □□ (□) es el *e-ésimo* elemento de la representación de alto nivel

$_{n}$□∈ R^{72} , que integra la información espacial, espectral y temporal de X_n .



F ig. 6 El bloque superior es la estructura de la LSTM bidireccional . Concatenamos las salidas de LSTMP y LSTMP como la salida del LSTM bidireccional, $Y \in R^{2T\times 72}$. El bloque central representa la proyección de las salidas del LSTM bidireccional. El bloque inferior denota la generación de la atención temporal.

Clasificador

Basándonos en la representación de alto nivel □□ de las

definirse del siguiente modo:

$$□□□ = □ □ □ □ □ □ □(□^{□} □_{□} + □)^{□} \quad (21)$$

donde W^p , b^p son parámetros aprendibles y $Pre \in R^C$

señales EEG, aplicamos una capa totalmente conectada y una activación *softmax*

denota la probabilidad de que □□ pertenezca a todas las clases □. En concreto, la clase de mayor probabilidad es la etiqueta predicha por 4D-aNN.

Experimento

En esta sección presentamos en primer lugar un conjunto de datos muy utilizado. A continuación, se describe la configuración del experimento. Por último, se presentan y discuten los resultados obtenidos.

Conjunto de datos SEED

El conjunto de datos SEED (Zheng y Lu, 2015) contiene 3 categorías diferentes de datos de emoción: positiva, neutra y negativa. Para cada tipo de emoción, se seleccionan 5 clips de película de unos 4 minutos de duración que pueden provocar la emoción deseada. En el experimento de recogida de señales EEG participan 15 sujetos sanos (7 hombres y 8 mujeres, con una edad de (23,27 ± 2,37)). Se realizan 3 grupos de experimentos para cada sujeto, y cada experimento consta de 15 procesos de visionado de clips. Cada proceso de visionado de clips puede dividirse en cuatro fases, que incluyen una pista de inicio de 5 segundos, un periodo de clip de 4 minutos, una autoevaluación de 45 segundos y un periodo de descanso de 15 segundos. El orden de las
15 clips se organiza de modo que no se muestren consecutivamente dos clips que provoquen la misma emoción. Las señales EEG de los experimentos se registran con un sistema ESI NeuroScan de 62 canales y se muestrean a 200 Hz. Además, las señales EEG seriamente contaminadas por electromiografía (EMG) y electrooculografía (EOG) se eliminan manualmente. A continuación, se aplica un filtro de paso de banda entre 0,3 y 50 Hz para filtrar el ruido.

Ajustes

El 4D-aNN propuesto toma un segmento 4D $X_n \in \mathbb{R}^{h \times w \times 2f \times 2T}$ como entrada. En este trabajo, adoptamos el mapa disperso 2D con h = 19 y w = 19 para mantener la relación posicional de los electrodos. Como se ha demostrado en trabajos anteriores, la combinación de las 5 bandas puede contribuir a obtener mejores resultados, de modo que fijamos
□ = 5. Para cada experimento, fijamos la longitud de los segmentos □ en 3, obteniendo unas 1128 muestras por experimento. A continuación, realizamos una validación cruzada quíntuple en cada experimento y calculamos la precisión de clasificación media (ACC) y la desviación estándar (STD) de 3 experimentos para cada sujeto. La media de ACC y STD de todos los sujetos se toma como el rendimiento final de nuestro método. Entrenamos la 4D-aNN en una GPU NVIDIA GTX 1080. Se aplica la optimización Adam para minimizar la función de pérdida. Fijamos la tasa de aprendizaje en 0,0003, el tamaño del lote en 12 y el máximo de épocas en 150.

Modelos de referencia

- HCNN (Li et al. 2018): Utiliza una arquitectura CNN jerárquica para el reconocimiento de emociones EEG, tomando mapas de características 2D DE extraídos de la banda γ como entradas. HCNN solo tiene en cuenta la información espacial de las señales de EEG.
- BiHDM (Li et al. 2020): Considera las diferencias asimétricas entre dos hemisferios para el reconocimiento de emociones EEG.
- RGNN (Zhong et al. 2020): Tiene en cuenta la topología biológica entre las diferentes regiones cerebrales para capturar las relaciones tanto globales como locales entre los diferentes canales de EEG.
- 4D-CRNN (Shen et al. 2020): Construye características DE extraídas de señales EEG en estructuras de características 4D y utiliza una red neuronal recurrente convolucional para extraer características espaciales, características espectrales y características temporales para el reconocimiento de emociones EEG.
- SST-EmotionNet (Jia et al. 2020): Utiliza una red de dos flujos para extraer características espaciales, espectrales y temporales. Además, SST-EmotionNet adopta los mecanismos de atención para mejorar su rendimiento en el reconocimiento de emociones EE G.

Resultados

Comparamos nuestro modelo con 5 modelos de referencia en el conjunto de datos SEED. La Tabla 1 presenta el promedio de ACC y STD de estos modelos para el reconocimiento de emociones EEG. HCNN utiliza la arquitectura CNN jerárquica para clasificar la emoción, pero sólo tiene en cuenta la información espacial de las señales EEG, alcanzando un 88,60% de precisión en la clasificación. BiHDM (Li et al. 2020) aplica cuatro RNN dirigidas para obtener una representación profunda de todas las señales de los electrodos del EEG, alcanzando una precisión de clasificación del 93,12%. RGNN tiene en cuenta la topología biológica entre las diferentes regiones del cerebro, alcanzando un 94,24% de precisión en la clasificación. precisión de clasificación. 4D -CRNN toma como entrada mapas de características 4D DE que contienen información espacial, espectral y temporal, y alcanza una precisión de clasificación del 94,74%. SST - EmotionNet utiliza una red de dos flujos con mecanismos de atención y alcanza una precisión de clasificación del 96,02%. Sin embargo, el tamaño de los datos de cada muestra de entrada de SST- EmotionNet es unas 4 veces mayor que el de 4D-aNN. En comparación con los modelos de referencia, la 4D-aNN propuesta alcanza el rendimiento más avanzado en el conjunto de datos SEED con división intra-sujeto. El ACC medio de todos los sujetos es del 96,10%. Los resultados de cada sujeto se muestran en la Fig. 7, y hay 9 sujetos (#5, #6, #8, #9, #10, #11, #12, #13 y #15) cuyos resultados son mejores que la media ACC. Específicamente, para hacer una comparación justa con 4D -CRNN, realizamos experimentos con 4D-aNN (DE) y 4D-aNN (PSD), que representan el 4D-aNN sólo toma características DE como entradas y sólo toma características PSD como entradas, respectivamente . La precisión de la 4D-aNN (DE) supera a la de la 4D-CRNN en un 0,65%, lo que indica la superioridad de la 4D -aNN propuesta. Cuando se compara con 4D-aNN (DE) y 4D-aNN (PSD), 4D- aNN muestra el mejor rendimiento, lo que indica la eficacia de la combinación de diferentes características.

T able 1 Rendimiento (media ACC y STD (%)) de los modelos comparados.

Modelo	SEMI LLA	
		ETS (%)
HCNN (Li et al. 2018)	88.60	2.60
BiHDM (Li et al. 2020)	93.12	6.06
RGNN (Zhong et al. 2020)	94.24	5.95
4D-CRNN (Shen et al. 2020)	94.74	2.32
SST-EmotionNet (Jia et al. 2020)	96.02	2.17
4D-aNN	96.10	2.61
4D-aNN (DE)	95.39	3.05
4D-aNN (PSD)	90.49	7.97

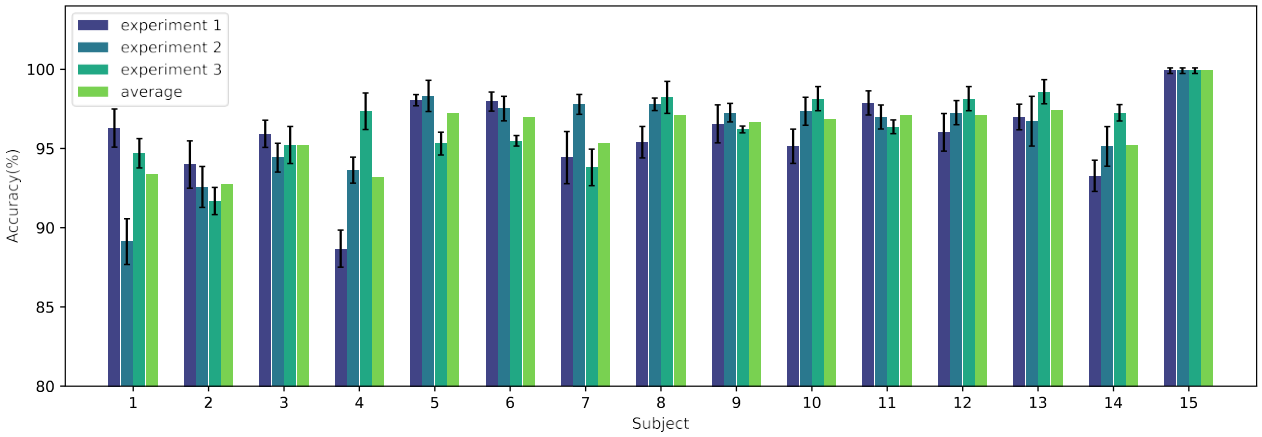
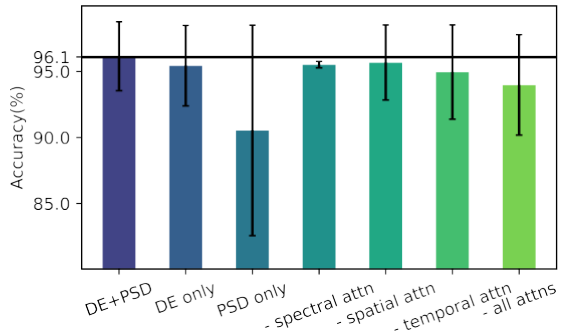


Fig. 7 Rendimiento de 4D-aNN en cada sujeto. En el conjunto de datos SEED, se realizan 3 experimentos para cada sujeto. Evaluamos el rendimiento de cada experimento y también presentamos la precisión de clasificación media para cada sujeto.

Para verificar la importancia de los mecanismos de atención en nuestro modelo, realizamos un experimento adicional de estudios de ablación en el conjunto de datos SEED. El experimento consiste en la ablación de los mecanismos de atención espacial, espectral y temporal. Evaluamos el rendimiento de 4D-aNN cuando se eliminan los mecanismos de atención espacial, espectral, temporal y todos los mecanismos de atención respectivamente. Como se muestra en la Fig. 8, cuando se elimina uno de los mecanismos de atención, la precisión de la clasificación disminuye. 4D -aNN sin el mecanismo de atención espectral disminuye un 0,63%, 4D-aNN sin el mecanismo de atención espacial disminuye un 0,47%, y 4D-aNN sin el mecanismo de atención temporal disminuye un 1,19%. En concreto, 4D -aNN sin todos los mecanismos de atención disminuye un 2,17%, que es el peor entre los modelos utilizados para la comparación. En conclusión, los resultados indican que los mecanismos de atención contribuyen al reconocimiento de emociones en EEG por su capacidad para captar patrones locales discriminatorios en los dominios espacial, espectral y temporal.



F ig. 8 Estudios de ablación en diferentes características de entrada y módulos de atención de 4D-aNN. "-" indica la ablación en determinados módulos de atención.

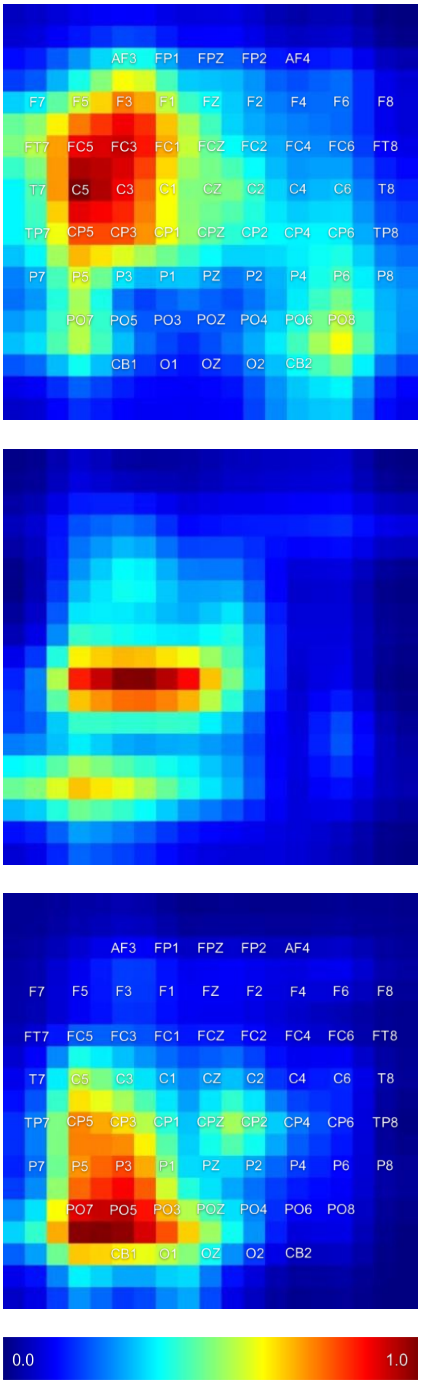
En particular, para explorar las regiones cerebrales críticas para diferentes emociones, representamos por separado los mapas térmicos de la actividad de los electrodos en la Fig. 9. Dibujamos los mapas de calor utilizando *Grad-* ++ (Chattopadhay et al. 2018), basándonos en los resultados experimentales del sujeto #15. ++ utiliza los mapas de características de la última capa convolucional y las puntuaciones de clase del clasificador para generar mapas de calor. Los mapas de calor son capaces de explicar qué regiones de entrada son importantes para las predicciones. En este trabajo, el tamaño de cada mapa térmico es de 19×19, que es el mismo que el mapa 2D disperso. Los elementos de los mapas de calor representan las contribuciones de las regiones cerebrales correspondientes al reconocimiento de las emociones objetivo. A partir de la Fig. 9, podemos observar las distintas distribuciones de las regiones cerebrales importantes con respecto a las diferentes emociones: los canales *FC5* , *FC3* y *C5* son importantes para el reconocimiento de las emociones positivas, los canales *CP5* , *CP3* y *CP1* son importantes para el reconocimiento de las emociones neutras, y los canales *PO7*, *PO5* y *P3* son importantes para el reconocimiento de la emoción negativa s para el sujeto #15. En particular,

las regiones cerebrales críticas podrían variar con diferentes sujetos , tiempo y emociones, de modo que los mecanismos de atención que permiten a 4D-aNN captar adaptativamente patrones discriminativos tienen sentido para el reconocimiento de emociones por EEG.

(a)

(b)

(c)



F ig. 9 Mapas de calor de la actividad de los electrodos basados en los resultados experimentales del sujeto nº 15. Las partes (a), (b) y (c) corresponden a emociones positivas, neutras y negativas, respectivamente. Las partes (a), (b) y (c) corresponden a emociones positivas, neutras y negativas, respectivamente. Las regiones de color rojo oscuro denotan contribuciones más significativas al reconocimiento de las emociones correspondientes.

Debate

Llevamos a cabo varios experimentos para investigar el uso de 4D-aNN que fusiona la información espacial-espectral-temporal y la eficacia de los mecanismos de atención en diferentes dominios para la clasificación de emociones EEG. En esta sección, discutimos tres puntos dignos de mención.

En primer lugar, para tratar la información espacial-espectral, aplicamos una CNN basada en la atención que consta de una red CNN, un módulo de atención espectral y un módulo de atención espacial. La red CNN extrae primero la representación espacial-espectral de las entradas. A continuación, el mecanismo de atención espectral se aplica a cada característica espectral para explorar la importancia de las distintas bandas de frecuencia y características. Además, el mecanismo de atención espacial se aplica a cada mapa de características 2 D para capturar de forma adaptativa las regiones cerebrales críticas. Las regiones cerebrales críticas y las bandas de frecuencia pueden variar en función de los individuos, las emociones y el tiempo, de modo que la capacidad de captar patrones discriminatorios de los módulos de atención mejora el rendimiento de la 4D-aNN.

En segundo lugar, para explorar las dependencias temporales en las representaciones 4D espacio-espectrales-temporales, utilizamos una LSTM bidireccional basada en la atención. La LSTM bidireccional extrae representaciones de alto nivel de las salidas de la CNN basada en la atención. A diferencia de los métodos tradicionales, que sólo utilizan la salida del último nodo de un LSTM para la clasificación u otras aplicaciones, con el mecanismo de atención temporal tenemos en cuenta las salidas de todos los nodos. El mecanismo de atención temporal asigna de forma adaptativa los pesos de los diferentes cortes temporales para que el contenido dinámico de las emociones en las representaciones 4D se pueda captar mejor.

En tercer lugar, para abordar la importancia de los mecanismos de atención, realizamos estudios de ablación en diferentes módulos de atención. 4D-aNN sin los mecanismos de atención espacial, espectral y temporal disminuye en un 0,47%, 0,63% y 1,19% la precisión de la clasificación, respectivamente. En particular, 4D-aNN sin todos los mecanismos de atención disminuye un 2,17%, que es el peor entre los modelos comparados. Los resultados experimentales demuestran la eficacia de los mecanismos de atención para captar de forma adaptativa patrones discriminativos.

Conclusión

En este artículo, proponemos el modelo 4D-aNN para el reconocimiento de emociones EEG. El modelo 4D-aNN toma como entrada representaciones 4D espacio-espectrales-temporales que contienen información espacial, espectral y temporal de las señales EEG. Integramos los mecanismos de atención en el módulo CNN y en el módulo LSTM bidireccional. El módulo CNN se ocupa de la información espacial y espectral de las señales EEG, mientras que los mecanismos de atención espacial y espectral capturan regiones cerebrales críticas y bandas de frecuencia de forma adaptativa. El módulo LSTM bidireccional extrae las dependencias temporales

en las salidas del módulo CNN, mientras que el mecanismo de atención temporal explora la importancia de diferentes cortes temporales. Los experimentos con el conjunto de datos SEED demuestran un mejor rendimiento que todas las líneas de base s. En particular, los estudios de ablación en diferentes módulos de atención muestran la eficacia de los mecanismos de atención en nuestro modelo para el reconocimiento de emociones EEG.

Referencia

Abbass SKGHA, Tan KC, Al -Mamun A, Thakor N, Bezerianos A, Li J (2018) Spatio -Spectral Representation Aprendizaje for Electroencephalographic Gait-Pattern Classification Ieee T Neur Sys Reh 26:1858-1867 doi:10.1109/TNSRE.2018.2864119

Alhagry S, Fahmy AA, El -Khoribi RA (2017) Emotion recognition based on EEG using LSTM recurrent neural network International Journal of Adva nced Computer Science and Applications 8:335 -358 doi:10.14569/IJACSA.2017.081046

Bamdad M, Zarshenas H, Auais MA (2015) Application of BCI systems in neurorehabilitation: a scoping review Disability and Rehabilitation: Assistive Technology 10:355-364 doi:10.3109/17483107.2014.961569

Blankertz B et al. (2016) La interfaz cerebro-ordenador de Berlín: avances más allá de la comunicación y el control Front Neurosci-Switz. 10:530 doi:10.3389/fnins.2016.00530

Brittona JC, Phan KL, Taylor SF, Welsh RC, Berridge KC, Liberzon I (2006) Neural correlates of social and nonsocial emotions: Un estudio fMRI Neuroimage 31:397-409 doi:10.1016/j.neuroimage.2005.11.027

Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-CAM++: Generalized Gradient - Based Visual Explanations for Deep Convolutional Networks (Explicaciones visuales basadas en gradientes generalizados para redes convolucionales profundas). Ponencia presentada en la 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12 -15 March 2018.

Craik A, He Y, Contreras-Vidal JL (2019) Aprendizaje profundo para tareas de clasificación de electroencefalogramas (EEG): una revisión J Neural Eng 16:031001 doi:10.1088/1741 - 2552/ab0ab5

Dolan RJ (2002) Emoción, cognición y comportamiento Science 298:1191-1194 doi:10.1126/science.1076358

Duan R-N, Zhu J-Y, Lu B-L (2013) Differential entropy feature for eeg-based emotion classification. Ponencia presentada en la 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), San Diego, CA, USA, 6-8 Nov. 2013.

Figueiredo GR, Ripka WL, Romaneli EFR, Ulbricht L (2019) Sesgo atencional para cara s emocionales en individuos deprimidos y no deprimidos: un estudio de seguimiento ocular. Ponencia presentada en la 2019 41ª Conferencia Internacional Anual de la IEEE Engineering in Medicine and Biology Society (EMBC), Berlín, Alemania, Alemania, 23-27 de julio de 2019.

Fiorinia L, Manci oppi G, Semeraro F, Fujita H, Cavallo F (2020) Unsupervised emotional state classification through physiological parameters for social robotics applications Knowledge -Based Systems 190 doi:10.1016/j.knosys.2019.105217

Jenke R, Peer A, Buss M (2014) Feature extraction and selection for emotion recognition from eeg IEEE Transactions on Affective Computing 5:327 -339 doi:10.1109/TAFFC.2014.2339834

Jia Z, Lin Y, Cai X, Chen H, Gou H, Wang J SST -EmotionNet: Espacial-Espectral-Temporal basado en atención 3D para el reconocimiento de emociones mediante EEG. En: Proceedingsof the28th ACMInternational Conference on Multimedia, Seattle, WA, USA, 2020. Association for Computing Machinery, pp 2909 - 2917. doi:10.1145/3394171.3413724

Jiaxin Ma, Tang H, Zheng W-L, Lu B-L Emotion recognition using multimodal residual LSTM network. En: Proceedings of the 27th ACM International Conference on Multimedia, Niza, Francia, 2019. Association for Computing Machinery, Nueva York, NY, USA, pp 176–183. doi:10.1145/3343031.3350871

Katsigiannis S, Ramzan N (2017) Dreamer: Una base de datos para el reconocimiento de emociones a través de señales de eeg y ecg de dispositivos inalámbricos de bajo coste disponibles en el mercado Ieee J Biomed Health 22:98-107 doi:10.1109/JBHI.2017.2688239

Kim M-K, Kim M, Oh E, Kim S -P (2013) A review on the computational methods for emotional state estimation from the human eeg Comput Math Method M 2013 doi:10.1155/2013/573734.

Krizhevsky A, Sutskever I, Hinton GE Clasificación de imágenes con redes neuronales convolucionales profundas. En: Advances in Neural Information Processing Systems, 2012. Curran Associates, Inc. pp 1097 -1105

Li J, Zhang Z, He H (2018) Redes neuronales convolucionales jerárquicas para el reconocimiento de emociones basado en EEG Cogn Comput 10:368-380 doi:10.1007/s12559-017-9533-x

Li M, Lu B-L (2009) Emotion classification ba sed on gamma-band EEG. Ponencia presentada en la Conferencia Internacional Anual 2009 de la Sociedad IEEE de Ingeniería en Medicina y Biología, Minneapolis, MN, EE.UU., 3-6 de septiembre de 2009.

Li Y et al. (2020) A Novel Bi -hemispheric Discrepancy Model for EEG Emotion Recognition Ieee T Cogn Dev Syst:1-1 doi:10.1109/TCDS.2020.2999337

Lotfia E, Akbarzadeh-T M-R (2014) Redes neuronales emocionales prácticas Redes neuronales 59:61-72 doi:10.1016/j.neunet.2014.06.012

Mühl C, Nijholt BAA, Chanel G (2014) A survey of affective brain computer interfaces: principles, state -of-the-art, and challenges Brain-Computer Interfaces 1:66-84 doi:10.1080/2326263X.2014.912881

Pfurtscheller G et al. (2010) La BCI híbrida Front Neurosci - Switz 4:3 doi:10.3389/fnpro.2010.00003

Shen F, Dai G, Lin G, Zhang J, Kong W, Zeng H (2020) EEG- based emotion recognition using 4D convolutional recurrent neural network Cogn Neurodynamics 14:815-828 doi:10.1007/s11571-020-09634-1

Shi L-C, Jiao Y-Y, Lu B-L (2013) Differential entropy feature for eeg-based vigilance estimation. Ponencia presentada en la 35ª Conferencia Internacional Anual 2013 de la IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japón, 3-7 de julio de 2013.

Song T, Zheng W, Song P, Cui Z (2020) EEG Emotion Recognition Using Dynamical Graph Conv olutional Neural Networks IEEE Transactions on Affective Computing 11:532-541 doi:10.1109/TAFFC.2018.2817622

Tao W, Li C, Song R, Cheng J, Liu Y, Wan F, Chen X (2020) EEG-based Emotion Recognition via Channel -wise Attention and Self Attention IEEE Transact ions on Affective Computing Afectiva:1 -1 doi:10.1109/TAFFC.2020.3025777

Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Módulo convolucional de atención por bloques. Computer Vision - ECCV 2018. Springer International Publishing, Cham. doi:10.1007/978 -3-030-01234- 2_1

Yan J, Zheng W, Xu Q, Lu G, Li H, Wang B (2016) Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech IEEE Transactions on Multimedia 18:1319 -1329 doi:10.1109/TMM.2016.2557721.

Yang Y, Wu Q, Fu Y, Chen X Red neuronal convolucional continua con entrada 3D para el reconocimiento de emociones basado en EEG. En: Cheng L, Leung ACS, Ozawa S (eds) Procesamiento neuronal de la información, 2018a. Springer International Publishing, pp 433 -433. doi:10.1007/978-3-030-04239-4_39

Yang Y, Wu QMJ, Zheng W-L, Lu B-L (2018b) Reconocimiento de emociones basado en EEG utilizando una red jerárquica con nodos de subred Ieee T Cogn Dev Syst 10:408-419 doi:10.1109/TCDS.2017.2685338

Zheng W-L, Lu B-L (2015) Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks IEEE Transactions on Autonomous Mental Development 7:162-175 doi:10.1109/TAMD.2015.2431497

Zheng W-L, Zhu J-Y, Lu B-L (2017) Identificación de patrones estables en el tiempo para el reconocimiento de emociones a partir de EEG IEEE Transactions on Affective Computing 10:417 -429 doi:10.1109/TAFFC.2017.2712143.

Zhong P, Wang D, Miao C (2020) EEG-Based Emotion Recognition Using Regularized Graph Neural Networks IEEE Transactions on Affective Computing:1-1 doi:10.1109/TAFFC.2020.2994159