Arquitectura de una red neuronal convolucional para el reconocimiento de expresiones faciales que representan emociones universales

Esmeralda Godinez Montero, Carlos J. Morales Hernandez, Marco A. Ramirez Hidalgo, Dr. Marco A. Moreno Armendáriz, Dr. Miguel S. Suárez Castañón
egodinezm1600@alumno.ipn.mx, cmoralesh1600@alumno.ipn.mx, mramirezh1600@alumno.ipn.mx

Escuela Superior de Cómputo,
Instituto Politécnico Nacional

Resumen

El presente artículo propone una arquitectura de red neuronal convolucional con mecanismos de atención llamada EmsaNet, la cual fue diseñada con el objetivo de reconocer las siete emociones universales propuestas por Paul Ekman, así como el estado neutral. También, se presenta el conjunto de datos DEM desarrollado para cumplir con la tarea de clasificación de emociones.

Introducción

A lo largo de los años, se ha buscado comprender las respuestas del ser humano ante estímulos recibidos en ciertas situaciones o eventos. A dichas respuestas se les ha definido como emociones; sin embargo, su estudio representa un gran reto debido a que son fenómenos subjetivos. Charles Darwin en su libro *The Expressions of Emotions in Man and the Animals* (1) señala que la expresión es universal en todos los seres humanos, mostrando que el ser humano tiene un origen común.

Basándose en las investigaciones de Charles Darwin y junto a sus propios experimentos, el psicólogo Paul desarrolló la teoría de las emociones universales, en donde identificó que, a pesar del trasfondo cultural y geográfico de las personas alrededor del mundo, todas manifiestan siete emociones universales (2): enojo, desprecio, disgusto, miedo, felicidad, tristeza y sorpresa.

Conocer las emociones de una persona con modelos automáticos conforma un área de tecnologías que pretenden entender los sentimientos, estados de ánimo e intenciones de las personas, lo que tiene aplicaciones prácticas en la prevención temprana de trastornos psicológicos, el mejoramiento de la comunicación entre una máquina y un humano, o personalización de experiencias en servicios.

En este artículo se presenta una arquitectura de red neuronal convolucional para reconocer las siete emociones universales más el estado neutral. Además, se introduce el *dataset* DEM, un conjunto de datos desarrollado en este trabajo para ayudar con la tarea de clasificación de emociones.

Estado del arte

Los primeros trabajos enfocados en la clasificación de emociones en imágenes sugieren dos pasos: la extracción de expresiones faciales y la detección de emociones. Primeramente, se identifican y extraen las características de las expresiones faciales; luego, se detectan las emociones universales a través de una clasificación de los resultados producidos en el primer paso.

Algunos trabajos que realizan dicha clasificación son (3), (4) y (5); sin embargo, estos enfrentaron dificultades al intentar clasificar emociones en imágenes con algunas variaciones, como aquellas donde el rostro se visualiza parcialmente o cuando la persona toca su cara con sus manos.

El *Large Scale Visual Recognition Challenge* es un concurso que evalúa algoritmos para la detección de objetos y la clasificación de imágenes. En 2012, la arquitectura ganadora del concurso fue AlexNet, una red neuronal convolucional. Esto demostró el gran potencial de este tipo de redes en el campo de visión por computadora, lo que llevó a que en trabajos posteriores se comenzara a utilizar redes neuronales convolucionales para el reconocimiento de expresiones faciales y la clasificación de emociones.

En (6) se presenta la arquitectura de una red neuronal convolucional y se demuestra que estas mejoran el rendimiento y la precisión en el reconocimiento de expresiones. Utilizando el *dataset* CK+, se logró una exactitud del 95,1 % ± 3,1 %.

Por otra parte, en 2021 se publicó (7), en donde se

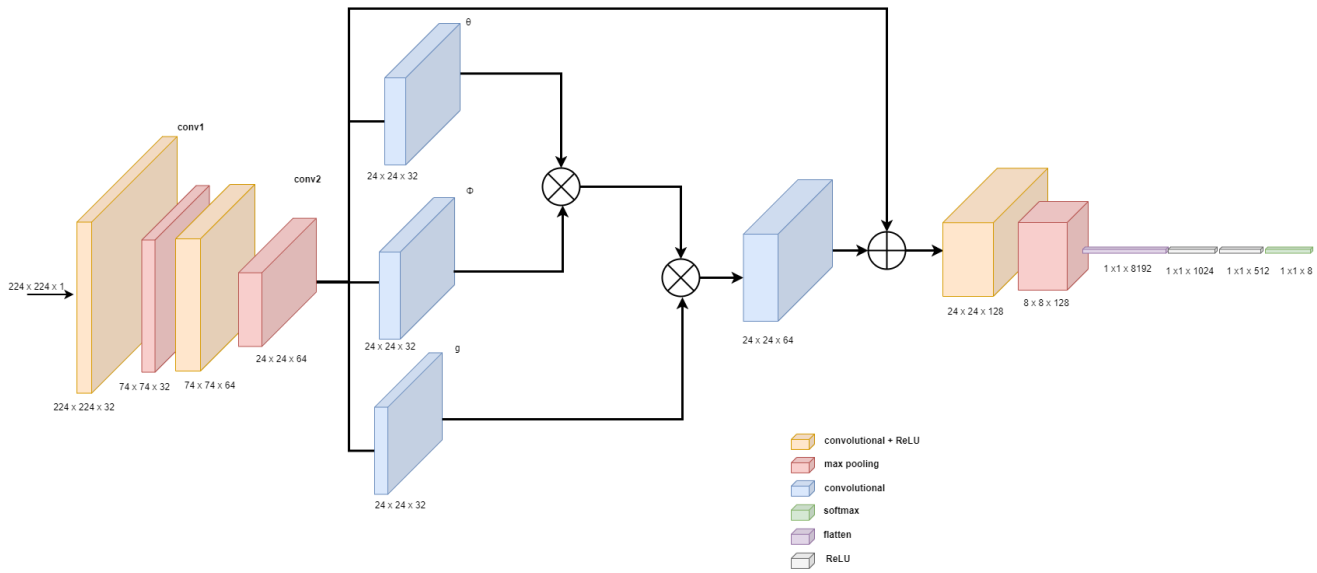

Figura 1: Arquitectura EmsaNet

propone una arquitectura de red neuronal usando capas convolucionales y mecanismos de atención que se enfocan en las partes del rostro que determinan la clasificación de emociones. Siendo así, se destaca que al experimentar y hacer pruebas con el *dataset* CK+, el algoritmo propuesto tuvo una exactitud del 98 %.

EmsaNet

Emotion Self-Attentive Convolutional Neural Network (EmsaNet), que se traduce al español como Red Neuronal Convolucional Auto-Atenta a las emociones, es la arquitectura propuesta en este trabajo, la cual se aprecia en la Figura 1.

La idea general fue el desarrollo de una red compuesta por capas convolucionales que incluyen un módulo de atención que, en específico, emplea mecanismos de autoatención conocidos como *self-attention blocks*. De tal forma que con las capas convolucionales se buscan características locales, mientras que con el módulo de atención *non-local* se detectan relaciones globales entres dichas características.

La red neuronal consiste en tres capas convolucionales con filtros de tamaño 5×5 y *strides* de una unidad. El número de filtros de cada capa fueron 32, 64 y 128, respectivamente, y como función de activación se usó ReLu (*Rectified Linear Unit*). Seguido de cada una de las tres capas convolucionales se colocaron capas *max pooling* con filtros de tamaño 3×3. En cuanto al mecanismo de atención, se usó un bloque *non-local* despues de la segunda capa de *max pooling*.

Finalmente, se incluyó una red completamente conectada para realizar la clasificación de las siete emociones y el estado neutral. En general, la capa de entrada cuenta con 1,024 neuronas, la capa oculta tiene 512 neuronas y la capa de salida ocho neuronas.

Dataset on Emotions among Mexicans

En este trabajo se desarrolló *Dataset on Emotions among Mexicans* (DEM), el cual es un conjunto de datos que se constituye de fotografías provenientes de personas posando las siete emociones universales descritas por Paul Ekman, y también se incluye un estado neutral que se puede interpretar como la ausencia de emociones.

En DEM participaron 246 voluntarios a los cuales se les tomó individualmente un máximo de ocho fotografías, cada foto correspondiente a una emoción y al estado neutral, recopilando así 1,941 imágenes. Los participantes son estudiantes o profesores de universidades y centros de investigación en la Ciudad de México.

La toma de fotografías fue realizada por los autores de este artículo y se recopilaron datos sobre los participantes, tales como: la escuela a la que pertenecían al momento de participar, edad y lugar de nacimiento.

Por otro lado, el etiquetado de DEM se llevó a cabo en dos etapas. La primera de ellas consistió en confirmar, por medio de cuatro anotadores externos, que la emoción que el participante pretendía expresar correspondía a la emoción que se le pidió realizar. La segunda etapa de etiquetado se efectuó con los resultados de la primera etapa y dos anotadores externos que fueron contratados para etiquetar todas las fotografías.

Como resultado, se obtuvo un etiquetado en el que participaron seis anotadores externos e independientes entre sí. Además, cabe resaltar que los autores de DEM intervinieron en esta etapa de etiquetado como anotadores.

Finalmente, el *dataset* DEM se compone de 1,557 fotografías con una resolución de 6,016 × 4,016 píxeles ordenadas por la emoción que representan, las cuales se seleccionaron al aplicar diversos criterios de elección, tomando en cuenta factores como la resolución en píxeles de las imágenes, la edad de los participantes y la opinión conjunta de los anotadores en el etiquetado.

En el Cuadro muestra la cantidad de imágenes por emoción en el *dataset* DEM, con mayor relavancia en fotos la emoción felicidad, mientras que el miedo contiene pocas imágenes.

En la Figura 2 se exponen cuatro ejemplares de fotografías contenidas en DEM. En la esquina superior izquierda, se aprecia una imagen de tristeza; en la esquina superior derecha, una de sorpresa; en la esquina inferior izquierda, una de disgusto; y en la esquina inferior derecha, una de enojo.

Figura 2: Imágenes contenidas en *Dataset on Emotions among Mexicans*.

Cuadro 1: Conteo de fotografías por clase en DEM.

Emoción	**Cantidad**
Desprecio	171
Disgusto	196
Enojo	171
Felicidad	258
Miedo	130
Neutral	285
Sorpresa	208
Tristeza	138
Total	**1,557**

Experimentos y resultados

Se llevaron a cabo un total de 110 experimentos con el objetivo de evaluar y comparar el rendimiento de diferentes arquitecturas de redes neuronales convolucionales en el reconocimiento de expresiones faciales que representan emociones universales.

Para ello, se utilizaron diversos conjuntos de datos, incluyendo CK+ con 357 fotografías, DEM con 1,557 imágenes, DISFA+ con 1,393 muestras, FER+ con 33,712 imágenes y una combinación de los cuatro conjuntos denominada EMOTION+.

Durante la etapa experimental, se emplearon cuatro arquitecturas de redes neuronales convolucionales ampliamente reconocidas en la literatura: Inception, VGG19, SqueezeNet y LeNet. Estas arquitecturas representan diferentes enfoques y niveles de complejidad en el diseño de las redes.

Cada arquitectura fue evaluada en los conjuntos de datos mencionados, realizando pruebas exhaustivas para medir su desempeño en el reconocimiento de expresiones faciales. Para cada experimento, se llevaron a cabo tareas de entrenamiento y evaluación utilizando técnicas de aprendizaje supervisado.

La métrica utilizada para comparar los resultados obtenidos en cada experimento fue la exactitud de las pruebas realizadas con cada conjunto de datos, de tal forma que se llevó un registro de dicha métrica en las evaluaciones de cada arquitectura.

En el Cuadro 2 se presentan los mejores resultados obtenidos en los experimentos. Es importante destacar que la arquitectura EmsaNet mostró un rendimiento sobre-

Cuadro 2: Mejores resultados de exactitud de las pruebas de todos los experimentos por arquitectura.

Experimento	VGG19	EmsaNet	LeNet	Inception	Squeeze
DEM *train* y CK+ *full*	0.7338	**0.7394**	0.6891	0.507	0.6331
CK+ *train* y CK+ *test*	0.863	**0.9041**	0.8219	0.6849	0.2857
DISFA+ *train* y DISFA+ *test*	**0.9714**	0.9678	0.9679	**0.9714**	0.8107
FER+ *train* y FER+ *test*	0.6637	0.7053	**0.7138**	0.5815	0.6881
EMOTION+ *train* y CK+ *portion*	0.9016	**0.9344**	0.8852	0.7049	0.8361
DEM a color *train* y DEM *test* a color	0.6058	**0.6282**	0.5984	0.6002	0.5702

saliente al ser la ganadora en cuatro de los seis mejores experimentos. Específicamente, se observa que EmsaNet obtuvo el mejor resultado en el experimento evaluado con el conjunto de datos EMOTION+ *train* y CK+ *portion*, alcanzando una exactitud de 0.9344.

A continuación, se presenta la matriz de confusión de uno de los experimentos (Figura 3), que revela la capacidad de la red neuronal para reconocer correctamente las emociones de desprecio, felicidad, tristeza, sorpresa y el estado neutral. Sin embargo, se observa que existe un porcentaje mínimo de confusión en la clasificación de las emociones de enojo y miedo.

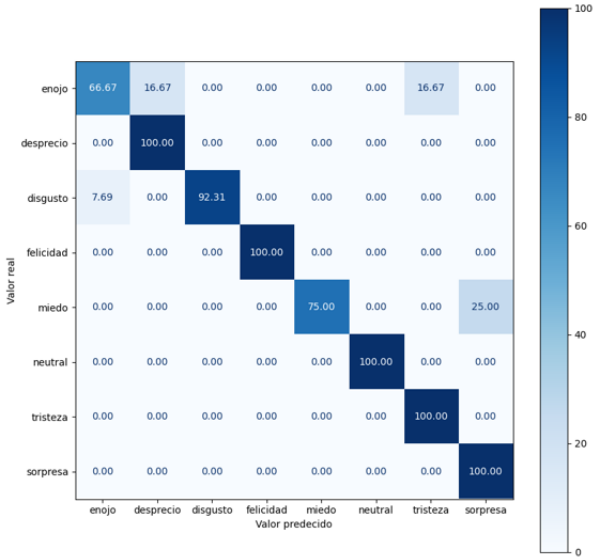

Figura 3: Matriz de confusión del entrenamiento de EmsaNet con EMOTION+ y evaluado con CK portion.

Por último, la Figura 4 muestra la representación visual del mecanismo de atención utilizado en la red EmsaNet. En esta figura, se utiliza un mapa de calor para resaltar las áreas a las que la red neuronal presta más atención durante la clasificación de las expresiones faciales.

En el mapa de calor, se puede observar que las regiones coloreadas en rojo indican las áreas de mayor atención por parte de la red neuronal. Estas áreas resaltadas corresponden principalmente a las cejas, los ojos y la boca, que son las regiones faciales más relevantes e informativas para el reconocimiento de emociones.

Figura 4: Resultados obtenidos del mecanismo de atención de la red EmsaNet.

Conclusiones y trabajo futuro

En este artículo se presentó EmsaNet, una arquitectura de red neuronal convolucional para reconocer emociones universales y el estado neutral. También, se introdujo el *dataset* DEM, el cual contiene fotografías de personas mexicanas posando las siete emociones universales de Paul Ekman y el estado neutral.

Con el propósito de analizar EmsaNet y el dataset DEM, se realizaron 110 experimentos para comparar los resul-

tados de todas las arquitecturas al hacer pruebas con los conjuntos de datos CK+, DISFA+ y FER+.

Después de la etapa de experimentación, se ha observado que el enfoque óptimo implica entrenar y evaluar los modelos de redes neuronales convolucionales con el mismo conjunto de datos, por lo que al realizar evaluaciones con otros conjuntos de datos, se han obtenido resultados deficientes. Así, se ha evidenciado que los conjuntos de datos CK+ y DISFA+ no son los más adecuados para el entrenamiento debido a sus limitaciones en términos de diversidad de personas y expresiones emocionales. Estas limitaciones impactan negativamente la capacidad de generalización del conocimiento en las arquitecturas de RNC.

Por otro lado, se ha identificado un desafío significativo en FER+ relacionado con el tamaño de sus fotografías, las cuales son tan pequeñas que los modelos no logran aprender correctamente este conjunto de datos. No obstante, la gran cantidad de fotos disponibles permite obtener buenos resultados en las etapas de entrenamiento y pruebas. En contraste, DEM se caracteriza por una amplia variedad de personas y emociones, lo cual dificulta el proceso de aprendizaje debido a la limitada cantidad de ejemplares fotográficos disponibles.

El conjunto de datos DEM fue etiquetado por un total de cinco anotadores, quienes presentaron un acuerdo bajo del 67 % en el número total de fotografías consideradas para cada una de las clases. Esta discrepancia tiene un impacto en el entrenamiento de los modelos y se refleja en los resultados de las pruebas, los cuales alcanzan un máximo del 62 %. Para mejorar estos resultados, se sugiere la incorporación de más anotadores durante el proceso de etiquetado. Además, se observó que las pruebas realizadas con DEM en formato a color solo presentaron una mejora del 1 %, lo cual resulta poco significativo.

Finalmente, se ha comprobado que el mecanismo de atención implementado en la red EmsaNet funciona de acuerdo con lo esperado, al otorgar una mayor importancia a las zonas más representativas para la manifestación de emociones, tales como las cejas, los ojos, la nariz y la boca.

La red EmsaNet destacó al obtener los mejores resultados en cuatro de los seis experimentos más relevantes. El rendimiento más sobresaliente se logró con una exactitud de prueba del 90.4 % al entrenar y evaluar con el conjunto de datos CK+. Asimismo, esta arquitectura demostró un rendimiento del 70.5 % al entrenar con FER+ y realizar pruebas con el mismo conjunto de datos, situándola al nivel del estado del arte.

Referencias

[1] C. Darwin, *The expression of the emotions in man and animals*. John Murray, 1872.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, pp. 124–129, 1971.

[3] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition based on facial components detection and hog features," *Scientific Cooperations International Workshops on Electrical and Computer Engineering Subfields*, pp. 803–816, 2014.

[4] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Elsevier*, 2009.

[5] J. Whitehill and C. W. Omlin, "Haar features for facs au recognition." [Online]. Available: https://inc.ucsd.edu/mplab/users/jake/haar.pdf

[6] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1510.02969

[7] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *CoRR*, 2019. [Online]. Available: http://arxiv.org/abs/1902.01019