LiExNet: MODELO PARA LA CLASIFICACIÓN DE EMOCIONES MEDIANTE EXPRESIONES FACIALES EN TIEMPO REAL

Jesus Joshua Muñoz Pacheco
Tecnológico Nacional de México
campus Chihuahua
Maestría en Ciencias de la ingeniería
Electrónica
Chihuahua, Chihuahua, México
jtec4550@gmail.com

***Resumen*— Actualmente, los modelos de reconocimiento de emociones a partir de expresiones faciales han evidenciado avances notables, favoreciendo la interacción humano-computadora en ámbitos como educación, salud y mercadeo. No obstante, muchas arquitecturas resultan demasiado complejas y costosas, dificultando su uso en tiempo real y en entornos con recursos limitados. En respuesta, este trabajo presenta LiExNet, un modelo ligero y eficiente inspirado en MobileNet, que emplea convoluciones separables en profundidad y un número reducido de parámetros. Entrenado principalmente con la base de datos CK+, LiExNet integra aumento de datos, normalización por lotes y ajuste dinámico de la tasa de aprendizaje, alcanzando resultados favorables en exactitud, F1-score, precisión y recall. Estos logros demuestran el potencial del modelo para implementaciones prácticas y de bajo costo computacional.**

Palabras clave*— *Reconocimiento de emociones, expresiones faciales, aprendizaje profundo, convoluciones separables, CK+

I. INTRODUCCIÓN

El reconocimiento automático de emociones a partir de expresiones faciales ha cobrado creciente relevancia en diversas áreas, como la interacción humano-computadora, el diagnóstico y seguimiento en entornos clínicos, la educación inteligente y la robótica social. Estas aplicaciones se basan en la premisa de que las expresiones faciales, en su mayoría, trascienden las barreras culturales y pueden ser asociadas a emociones básicas universales [1]. De acuerdo con Ekman y Friesen [1], emociones como la ira, el desprecio, el disgusto, el miedo, la felicidad, la tristeza y la sorpresa muestran patrones expresivos faciales reconocibles a través de distintas culturas, lo que sienta las bases para el desarrollo de sistemas computarizados de detección y clasificación emocional, lo que a su vez puede dar paso a la creación de sistemas de computación afectiva.

En años recientes, la evolución de las arquitecturas de aprendizaje profundo ha impulsado el desarrollo de modelos más precisos y eficientes. Se han desarrollado arquitecturas de modelos que emplean una combinación de redes, como el uso de redes 3D-CNN y ConvLSTM para la clasificación de expresiones usando conjuntos de datos como Extended Cohn-Kanade (CK+), alcanzando precisiones altas por arriba del 95% [2]. Pero además de ello existen otros trabajos que proponen un enfoque basado en otras arquitecturas como LeNet, combinando diversos conjuntos de datos incluyendo la base de datos Japanese Female Facial Expression (JAFFE) para reconocer emociones en tiempo real, reportando altas precisiones en el entrenamiento y validación de los modelos [3]. Estas investigaciones evidencian el avance en la precisión y la capacidad de procesamiento en tiempo real, requisitos indispensables para aplicaciones prácticas e interactivas.

Sin embargo, a pesar de los avances logrados, aún persisten desafíos que limitan la escalabilidad y aplicabilidad generalizada de estas soluciones. Esto se hace evidente en revisiones sistemáticas [4], donde se identifican problemas como la limitada capacidad de generalización y adaptabilidad de los modelos actuales, así como la necesidad de mayor confianza en sus predicciones. Además, aspectos como la variabilidad en condiciones de iluminación, diversidad cultural y factores contextuales complejos dificultan la robustez de los sistemas en situaciones del mundo real.

En este contexto, el presente artículo introduce LiExNet, un modelo diseñado para la clasificación de emociones mediante expresiones faciales en tiempo real, orientado a su implementación en sistemas de bajos recursos que puedan ser usados en sistemas de control como, por ejemplo, sistemas de computación afectiva. En particular, LiExNet adopta una arquitectura de aprendizaje profundo basada en convoluciones separables (Depthwise Separable Convolutions) y una reducción intencional en la cantidad de filtros por capa, buscando así una mayor eficiencia computacional sin sacrificar la capacidad representativa. Adicionalmente, el modelo utiliza técnicas de normalización por lotes regularización L2, y una estrategia de Dropout ajustada para mitigar el sobreajuste. Esta combinación de técnicas está orientada a obtener un sistema capaz de ofrecer predicciones estables en tiempo real, con un balance adecuado entre exactitud, eficiencia y adaptabilidad.

El resto del artículo se organiza de la siguiente forma: en la Sección II se describe la metodología y la arquitectura propuesta; en la Sección III se presenta el desarrollo del modelo y su sustentabilidad matemática; en la Sección IV se discuten las implicaciones y limitaciones del enfoque el entorno de evaluación y los resultados obtenidos, así como las observaciones finales y proyecciones futuras.

II. METODOLOGÍA

A. Conjuntos de datos

Para el entrenamiento y evaluación del modelo se emplearon las bases de datos CK+ y JAFFE. CK+ es una de las más utilizadas en el reconocimiento de expresiones faciales, ya que contiene secuencias de imágenes de alta variabilidad emocional. Por otro lado, JAFFE ofrece un

conjunto más limitado de imágenes, todas provenientes de un grupo reducido de sujetos, lo que dificulta la generalización y el aprendizaje profundo cuando se usa en solitario. Durante esta investigación, las pruebas más relevantes se obtuvieron a partir de CK+, dado que al entrenar con JAFFE el modelo no logró un aprendizaje adecuado debido a la escasa diversidad de patrones faciales. Esta limitación con JAFFE se considera un reto a abordar en trabajos futuros (ver Sección IV).

B. Preprocesamiento de imágenes

Las imágenes se convirtieron inicialmente a escala de grises y luego a formato RGB, para mantener la consistencia de las capas convolucionales estándares en redes de visión por computadora. Posteriormente, se ajustó su tamaño a 128×128 píxeles. La elección de esta resolución surgió tras experimentaciones preliminares, sin aplicar técnicas adicionales de normalización, más allá de la división de los valores de intensidad por 255.0. El objetivo principal fue conservar un balance entre detalle facial y eficiencia de procesamiento. Además, dadas las condiciones relativamente controladas de CK+, no se requirieron correcciones adicionales de iluminación o pose.

C. Aumento de datos

Para mejorar la generalización del modelo y robustecer el entrenamiento ante variaciones leves en la apariencia facial, se aplicaron transformaciones geométricas como desplazamientos horizontales y verticales, aumento y volteo horizontal. Estas transformaciones fueron seleccionadas por su bajo impacto en la correcta identificación de patrones faciales, descartando, por ejemplo, el volteo vertical, que podría alterar las características distintivas del rostro. El aumento de datos se aplicó exclusivamente sobre el conjunto de entrenamiento.

D. Arquitectura del modelo

LiExNet se diseñó tomando como referencia la filosofía de MobileNet, en la cual se emplean convoluciones separables en profundidad para reducir significativamente el número de parámetros y operaciones, manteniendo una capacidad representativa adecuada. Previamente, se consideraron arquitecturas populares como MobileNet, SqueezeNet, VGG16 y DenseNet121, con las cuales se obtuvieron buenos resultados, pero con un elevado número de parámetros. Un primer intento híbrido, combinando bloques de dichas arquitecturas, superaba el medio millón de parámetros y no ofrecía una mejora sustancial en eficiencia. En contraste, LiExNet presenta aproximadamente 7,471 parámetros, lo que resulta en una mayor velocidad de inferencia y un menor costo computacional.

La arquitectura resultante incluye capas convolucionales separables, normalización por lotes y capas de pooling para la reducción progresiva de dimensiones. Además, se introdujo regularización L2 y una capa Dropout (0.3) en las etapas finales, mitigando el sobreajuste sin impactar drásticamente en el desempeño. Puede verse la arquitectura más detalladamente en la Figura 1.

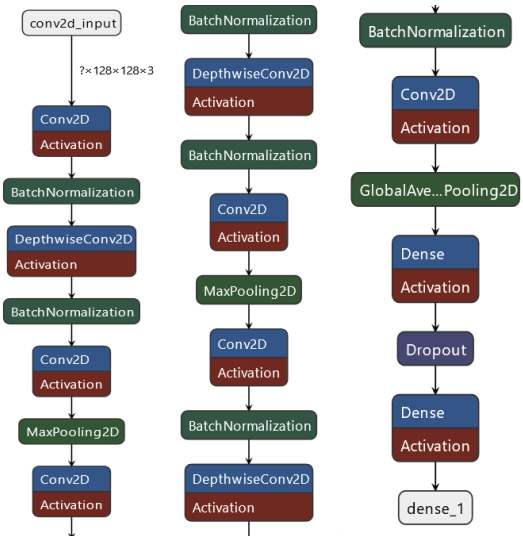

Figura 1. Arquitectura de la red LiExNet

E. Entrenamiento

Se utilizó el optimizador Adam con una tasa de aprendizaje inicial de 1e-4. El entrenamiento se llevó a cabo durante 600 épocas, con lotes de 64 imágenes. Para adaptar dinámicamente el ritmo de aprendizaje, se empleó el callback ReduceLROnPlateau, mientras que un mecanismo de Early Stopping detuvo el entrenamiento cuando la pérdida de validación dejó de mejorar por un número determinado de épocas. El entrenamiento se realizó sobre un CPU AMD Ryzen 7 7700, completándose en aproximadamente 35 minutos.

F. Evaluación y métricas de desempeño

El conjunto de datos se dividió en entrenamiento y validación, sin un conjunto de prueba separado. Además de la exactitud (Accuracy), se calculó el F1-score, la Precisión y el Recall, métricas comunes en la literatura que permiten valorar el rendimiento del modelo de forma más integral. También se empleó una matriz de confusión para analizar la distribución de errores entre las distintas emociones.

En la Sección IV discute los resultados obtenidos y su relevancia, incluyendo comparaciones con otros modelos reportados en la literatura.

III. DESARROLLO DEL MODELO Y FUNDAMENTOS MATEMÁTICOS

La arquitectura de LiExNet se sustenta en principios de eficiencia y simplicidad, inspirados en modelos ligeros como MobileNet. Para lograrlo, se adoptan convoluciones separables en profundidad, las cuales reducen significativamente el número de parámetros y operaciones requeridas, facilitando la ejecución en tiempo real y en entornos con recursos de cómputo limitados.

A. Convoluciones estándar

Una convolución bidimensional tradicional se puede describir de la siguiente manera: dada una entrada $\boldsymbol{X} \in R^{H\times W\times D_{in}}$, donde H y W son el alto y ancho de la imagen y D_{in} es la profundidad o numero de canales, y un conjunto de filtros $\boldsymbol{W} \in R^{k\times k\times D_{in}\times D_{out}}$, la operación convolucional para obtener la salida Y es:

$$Y_{h,w,d_{out}} = \sum_{d_{in}=1}^{D_{in}} \sum_{u=1}^{k} \sum_{v=1}^{k} X_{h+u-1,w+v-1,d_{in}} \cdot W_{u,v,d_{in},d_{out}}$$

donde k es el tamaño del kernel. Esta operación mezcla espacial y canal de manera simultánea, resultando en un costo computacional de $O(HWD_{in}D_{out}k^2)$

B. *Convoluciones separables en profundidad*

En contraste, las Depthwise Separable Convolutions descomponen la convolución estándar en dos etapas: una convolución por canal seguida de una convolución 1×1 pointwise convolution. Esta separación reduce drásticamente el número de operaciones.

1) *Convolucion depthwise*

Se aplica un filtro separado por cada canal de entrada. Si $\boldsymbol{W_{dw}} \in R^{k \times k \times D_{in}}$ representa el conjunto de filtros depthwise (uno por canal), la salida intermedia **Z** se calcula como:

$$Z_{h,w,d_{in}} = \sum_{u=1}^{k} \sum_{v=1}^{k} X_{h+u-1,w+v-1,d_{in}} \cdot W_{dw,u,v,d_{in}}$$

Esta fase opera únicamente sobre la dimensión espacial de cada canal por separado, con un costo de $O(HWD_{in}k^2)$.

2) *Convolucion pointwise*

Es Posteriormente, se combina la información resultante aplicando una convolución 1×1, que mezcla canales sin alterar el espacio. Con $\boldsymbol{W_{pw}} \in R^{1 \times 1 \times D_{in} \times D_{out}}$:

$$Y_{h,w,d_{out}} = \sum_{d_{in}=1}^{D_{in}} Z_{h,w,d_{in}} \cdot W_{pw,1,1,d_{in},d_{out}}$$

El costo aquí es de $O(HWD_{in}D_{out})$

Al combinar ambas etapas, el costo total es aproximadamente:

$$O(HWD_{in}k^2 + HWD_{in}D_{out})$$

Comparado con el costo original la reducción es significativa cuando D_{in} y D_{out} son relativamente grandes, mejorando la eficiencia sin sacrificar la capacidad representativa.

C. *Beneficios para el reconocimiento de emociones*

La reducción de parámetros y operaciones es especialmente relevante en el reconocimiento de emociones, ya que le permite a LiExNet operar a velocidades elevadas, respondiendo con baja latencia a cambios en las expresiones faciales del usuario. Además, con su arquitectura ligera es más fácil de implementar en plataformas embebidas, como la NVIDIA Jetson TX2, donde la capacidad de cómputo y la disponibilidad de memoria son restringidas y aunque el objetivo principal de las Depthwise Separable Convolutions es la eficiencia, su utilización combinada con normalización por lotes, regularización L2, y técnicas de aumento de datos permite que el modelo mantenga un alto rendimiento, sin incurrir en un sobreajuste excesivo ni perder capacidad para distinguir sutilezas en las expresiones faciales.

Este desarrollo matemático y conceptual sienta las bases de la arquitectura LiExNet, cuyos resultados experimentales se presentarán en la siguiente sección, evidenciando su eficacia y pertinencia para la tarea de reconocimiento de emociones en tiempo real.

IV. Resultados y Discusión

La evaluación del desempeño de LiExNet se llevó a cabo utilizando el conjunto de validación de la base de datos CK+. El modelo alcanzó una exactitud (Accuracy) cercana al 96.56%, reflejando su capacidad para clasificar correctamente la mayoría de las expresiones emocionales presentadas. Asimismo, se obtuvieron métricas complementarias que permiten apreciar la calidad de la clasificación desde distintas perspectivas: un F1-score promedio ponderado de 0.90, una Precisión (P) de 0.91, y un Recall (R) de 0.90. Estos valores evidencian un desempeño sólido en la identificación de cada clase emocional, incluso considerando posibles desbalances en la distribución de estas. Estas métricas pueden verse de manera más completa en las figuras 1, 2, 3 y 4.

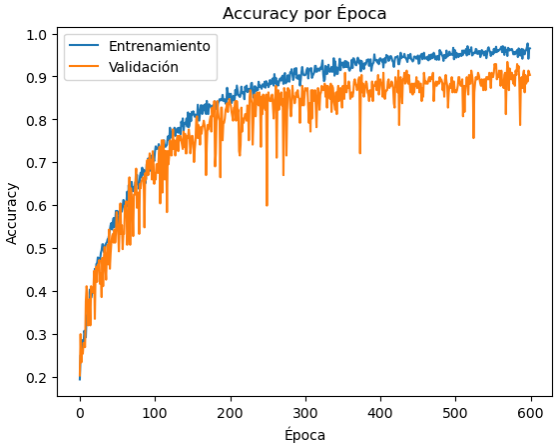

Figura 2. Grafica de Accuracy de la red LiExNet

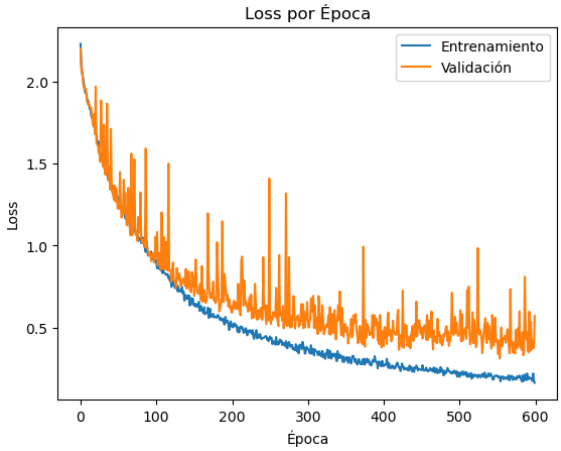

Figura 3. Grafica de perdida de la red LiExNet

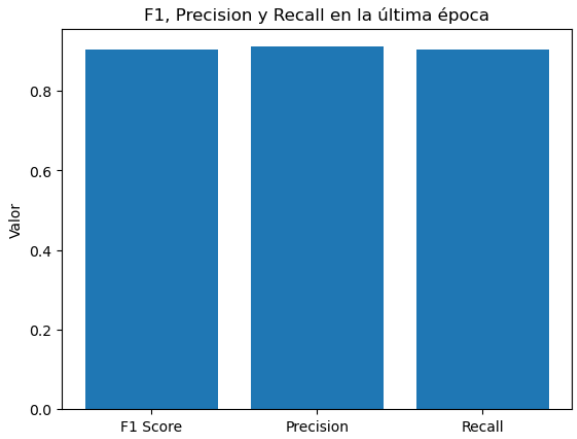

Figura 3. Grafica de barras de las métricas F1, Precisión y Recall de la última época de la red.

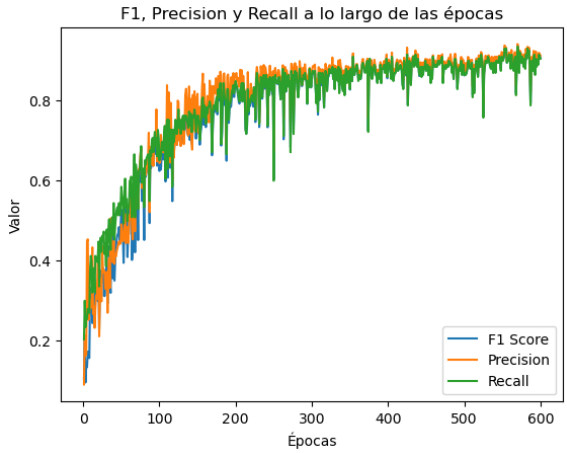

Figura 4. Grafica de barras de las métricas F1, Precisión y Recall atravez de las épocas de entrenamiento de la red.

La Fig. 5 presenta una matriz de confusión que ilustra la distribución de los aciertos y errores entre las siete emociones evaluadas. Si bien la mayoría de las categorías muestran altos valores diagonales, evidenciando un buen reconocimiento, es posible apreciar ligeras confusiones entre ciertas expresiones con rasgos visuales similares. Estos hallazgos sugieren que, si bien LiExNet es robusto ante la variabilidad presente en CK+, el modelo podría beneficiarse de una mayor diversidad en los datos de entrenamiento para afinar su habilidad de discriminar matices sutiles en algunas emociones, esto puede verse también en el pequeño sobreajuste que sufre la red en los momentos finales de su entrenamiento que aunque sea reducir marca un problema a resolver como parte del trabajo a futuro.

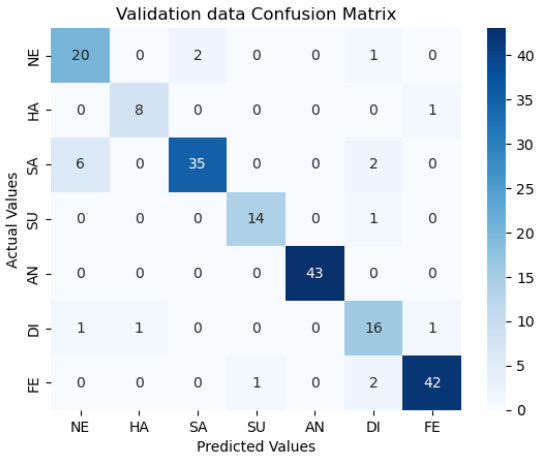

Figura 5. Matriz de confusión de la red LiExNet

Además de las métricas numéricas, se evaluó la capacidad de LiExNet para operar en tiempo real. Al implementar el modelo en una plataforma NVIDIA Jetson TX2 y procesar secuencias de video MP4, se alcanzaron tasas de inferencia suficientemente altas para brindar retroalimentación inmediata al usuario. Esta capacidad de inferencia rápida, combinada con su reducido número de parámetros (aprox. 7,471), posiciona a LiExNet como una solución atractiva para aplicaciones embebidas las métricas de dicha implementación pueden verse en la Tabla 1.

Tabla 1. Métricas y recursos de la implementación de la red en el sistema NVIDIA Jetson TX2.

Métrica	Valor
Fotogramas por segundo (FPS)	530.70
Uso de CPU	75.48%
Memoria utilizada (GB)	3.5/7.8

En términos comparativos, el rendimiento de LiExNet se ubica dentro del rango de las soluciones reportadas entre 2014 y 2023, según lo revisado en la literatura. Si bien modelos más complejos pueden alcanzar niveles ligeramente superiores de exactitud, estos suelen requerir un mayor consumo de recursos y tiempos de inferencia menos adecuados para entornos en tiempo real. La propuesta presentada en este articulo busca equilibrar la precisión con la eficiencia, obteniendo así un sistema confiable y con un costo computacional significativamente más bajo que las arquitecturas tradicionales de alto desempeño (por ejemplo, VGG16, DenseNet121, o combinaciones híbridas previas ensayadas). Puede verse la comparación de LiExNet en cuanto a su precisión con otros modelos que utilizan la misma base de datos [4] en la Figura 6.

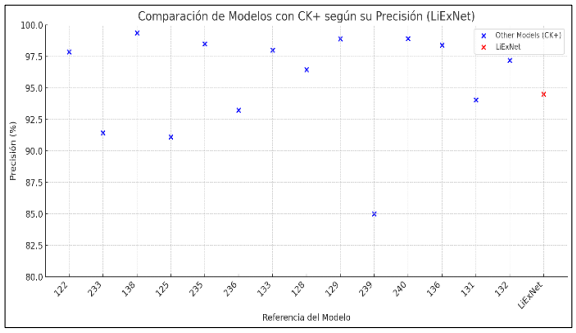

Figura 6. Comparación de la red LiExNet junto con las referencias de [4].

En suma, los resultados confirman la viabilidad de LiExNet como una alternativa eficiente para el reconocimiento de emociones en tiempo real. Si bien el modelo destaca por su desempeño y ligereza, el análisis de errores sugiere que la unión de las bases de datos JAFFE y CK+, podrían mejorar aún más su capacidad de generalización. Estos aspectos serán explorados en trabajos futuros, donde se prevé la integración de estas bases de datos para potenciar aún más la adaptabilidad del modelo.

AGRADECIMIENTOS

A mi laptop quien entreno variaciones del mismo modelo durante semanas, a mi carro que me trajo todos los días a trabajar sin cansarse, a los sábados por ser día de One Piece y darme ánimos de continuar, al profesor Abimael quien me ayudo en mi proceso de aprendizaje más de lo que otros profesores lo hicieron (no le diga a Juan), a nuestro señor Cthulhu quien vive por siempre en nuestros corazones, salve usted poderoso señor del cosmos.

REFERENCIAS

[1] P. Ekman y W. V. Friesen, "Constants across cultures in the face and emotion," Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124–129, 1971.

[2] R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra y S. Singh, “Facial expression recognition in videos using hybrid CNN & ConvLSTM”, Int. J. Inf. Technol., marzo de 2023. [En línea]. Disponible: https://doi.org/10.1007/s41870-023-01183-0 [Accedido: 9 de septiembre de 2024]

[3] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh y A. Akan, "Real Time Emotion Recognition from Facial Expressions Using CNN Architecture," en 2019 Medical Technologies National Congress (TIPTEKNO), Izmir, Turquía, 2019, pp. 1–4. DOI: 10.1109/TIPTEKNO.2019.8971998.

[4] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi y U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," Information Fusion, vol. 102, pp. 102019, 2024. DOI: 10.1016/j.inffus.2023.102019.

[5] J. Almeida y F. Rodrigues, "Stress detection using facial expressions and deep learning: A real-time approach with fine-tuned models," Journal of AI Research, vol. 15, pp. 234–245, 2024. DOI: 10.1016/j.aires.2024.03.004.

ANEXOS

Se deja a continuación el repositorio donde obtener los códigos de implementación y de entrenamiento de la red, así como cualquier tipo de material adicional disponible.

https://github.com/KorakiMx/LiExNet