Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

"Yo, como integrante de la comunidad estudiantil del Tecnológico de Monterrey, soy consciente de que la trampa y el engaño afectan mi dignidad como persona, mi aprendizaje y mi formación, por ello me comprometo a actuar honestamente, respetar y dar crédito al valor y esfuerzo con el que se elaboran las ideas propias, las de los compañeros y de los autores, así como asumir mi responsabilidad en la construcción de un ambiente de aprendizaje justo y confiable"

"Inteligencia artificial avanzada para la ciencia de datos II"

Momento de Retroalimentación:

Reto Evaluación

Equipo:

Frida Cano Falcón A01752953

Jorge Javier Sosa Briseño A01749489

Guillermo Romeo Cepeda Medina A01284015

Daniel Saldaña Rodríguez A00829752

Fecha de entrega: 15 de noviembre de 2023

Introducción

En el siguiente trabajo se realizó una investigación de 4 trabajos de publicación en revistas de investigación relacionadas a la visión computacional, los trabajos a investigar tienen en común el desarrollo de modelos de inteligencia artificial para la detección de rostros, poses y características, asi como reconocimiento facial y otras áreas de la visión computacional. El propósito de esta investigación más allá de revisar la utilidad de dichos modelos es revisar las métricas utilizadas para medir el desempeño de los mismos, así como también realizar o revisar las comparaciones de dichos modelos con otros utilizando dichas métricas de desempeño. Esto con el propósito de incluir las métricas que consideremos de mayor relevancia en nuestro reto para medir el desempeño de los modelos que estamos utilizando.

DeepMatcher: A deep transformer-based network for robust and accurate local feature matching

En el mismo, se presenta una nueva arquitectura para realizar el emparejamiento de características locales en imágenes llamada DeepMatcher, la cual está basada en la red de transformers. A lo largo del trabajo se muestra con detenimiento el funcionamiento de la misma, pero en este resumen nos enfocaremos en la métrica de desempeño que utilizaron para hacer comparaciones entre modelos. En el artículo se realizan varias comparaciones de diferentes datasets y muchos modelos, pero las métricas son las mismas, a continuación se describe la evaluación de estimación de poses en interiores.

Dataset:

Se utiliza el dataset ScanNet, que consta de 1513 secuencias con imágenes RGB y poses de referencia en entornos interiores, . Se seleccionan 230 millones de pares de imágenes (todas se reducen a un tamaño de 640x480.) con valores de superposición (overlap) entre 0.4 y 0.8 como conjunto de entrenamiento, y se reserva un conjunto de prueba.

Métrica de rendimiento:

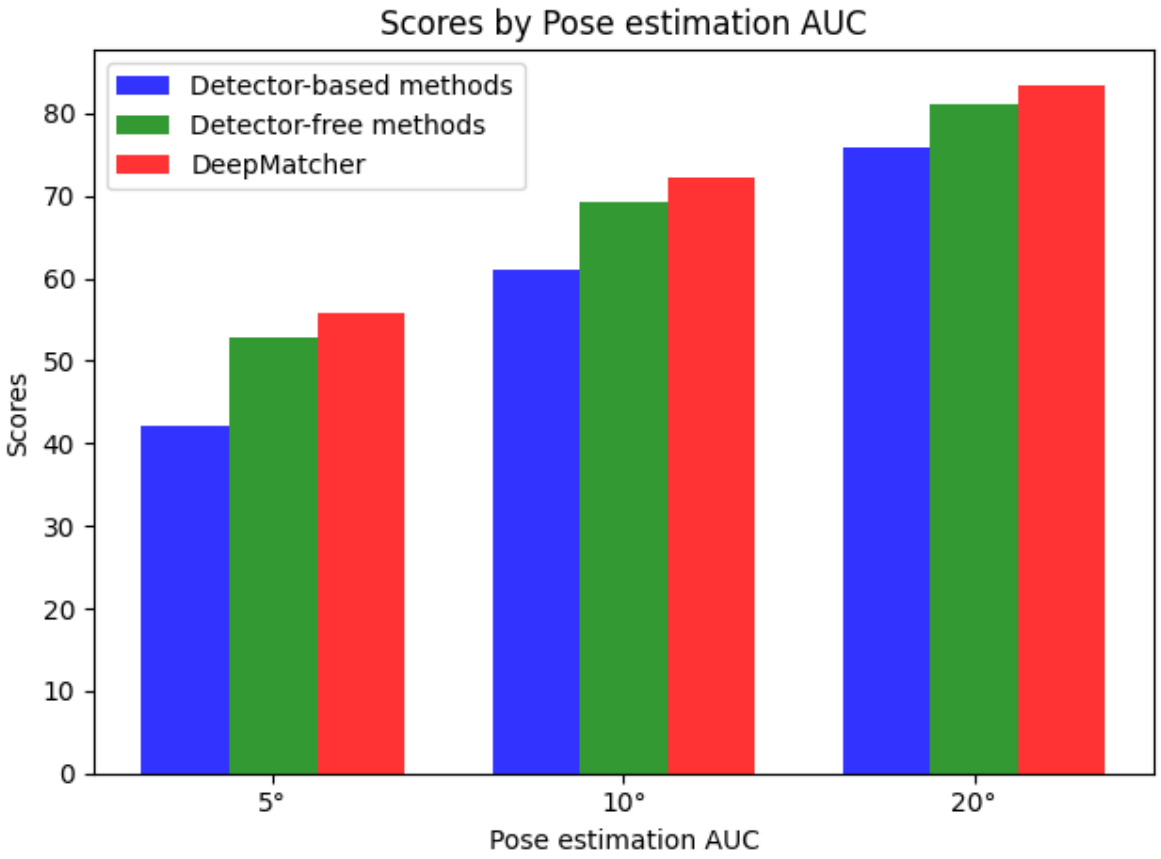
La métrica principal es el área bajo la curva (AUC) de los errores de poses a umbrales específicos. Los errores de poses se definen como las máximas diferencias traslacionales y

rotacionales entre las poses de referencia y las poses predichas por la red. Se utilizan tres límites de error de AUC@(5,10,20). Dados los errores acomulativos de las áreas se realiza una métrica de precisión para realizar la comparación entre modelos. Una mayor descripción del funcionamiento de la métrica de rendimiento puede ser encontrada en el artículo. Otro elemento que se utiliza como métrica de rendimiento es el número de consumo de GFLOPs, (operaciones de punto flotante por segundo).

Métrica de rendimiento:

	Matcher	**Pose estimation AUC**		
Empty Cell	Empty Cell	**@5°**	**@10°**	**@20°**
Detector-based methods				
SuperPoint (DeTone et al., 2018a)	SuperGlue (Sarlin et al., 2020)	42.18	61.16	75.96
	DenseGAP (Kuang et al., 2021)	41.17	56.87	70.22
	ClusterGNN (Shi et al., 2022)	44.19	58.54	70.33
Detector-free methods				
–	DRC-Net (Li et al., 2020)	27.01	42.96	58.31
	LoFTR (Sun et al., 2021)	52.80	69.19	81.18
	QuadTree (Tang et al., 2022)	54.60	70.50	82.20
	TopicFM (Truong Giang et al., 2022)	54.10	70.10	81.60
	MatchFormer (Wang et al., 2022)	52.91	69.74	82.00
	ASpanFormer	55.30	71.50	83.10

	(Chen, Luo et al., 2022)			
	DeepMatcher	55.71	72.25	83.49

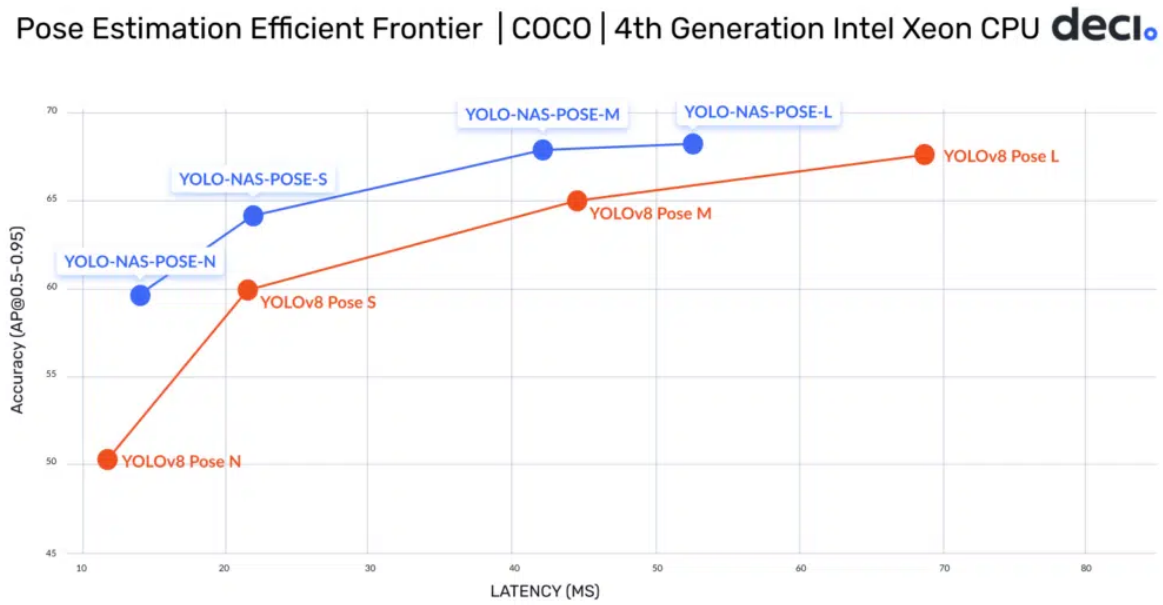


Los métodos detector-free muestran un rendimiento superior en comparación con los métodos basados en detectores, especialmente en situaciones con cambios significativos de punto de vista. DeepMatcher y DeepMatcher-L superan considerablemente a los métodos de última generación, tanto basados en detectores como libres de detectores, demostrando su eficacia en la tarea de estimación de poses en interiores.

Nuestro modelo de detección de pose puede ser re entrenado con un dataset clasificado y se puede hacer la comparación de imágenes para obtener la métrica de AUC y tener comparaciones entre la precisión de diferentes iteraciones del modelo, busque en la documentación de Ultralytics pero por ningún lado ellos muestran algún porcentaje de precisión sobre algun dataset, sin embargo encontré un articulo en donde se registra la precisión de los diferentes modelos de yolov8 pose, encontrando el YOLOv8 POse L como el

de mejor precisión (el testing se realizo con el dataset de COCO). En este artículo se utilizó una función de pérdida para obtener la métrica de precisión.

	Number of Parameters (In millions)	**AP@0.5-0.95**	**Latency (ms) Intel Xeon gen 4th (OpenVino)**	**Latency (ms) Jetson Xavier NX (TensorRT)**	**Latency (ms) NVIDIA T4 GPU (TensorRT)**
YOLO-NAS N	9.9M	59.68	14	15.99	2.35
YOLO-NAS S	22.2M	64.15	21.87	21.01	3.29
YOLO-NAS M	58.2M	67.87	42.03	38.40	6.87
YOLO-NAS L	79.4M	68.24	52.56	49.34	8.86



An attention-based CNN for automatic whole-body postural assessment

En este estudio se propone un enfoque basado en una red neuronal convolucional para realizar una evaluación postural. La red está diseñada para procesar imágenes a color, y predice los puntajes REBA (Rapid Entire Body Assessment) de seis partes del cuerpo.

La estructura de la red consta de una rama base que extrae las características globales, un módulo transformador espacial para la predicción de múltiples regiones de atención, y un número correspondiente de ramas para la extracción detallada de características locales.

La evaluación del método al usar el conjunto de datos pH36M supera a los métodos de referencia, esto demuestra la eficacia de la red propuesta.

Métricas de rendimiento

Se emplean dos métricas de rendimiento. Se utiliza el error absoluto medio (MAE) y el coeficiente Kappa de Cohen. No se usó la precisión como métrica de rendimiento ya que no se considera confiable para distribuciones desbalanceadas de riesgo.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i^* - y_i|$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Resultados

Results on pH3.6M (front-view). In "Ours-x", x refers to the number of branches.

MAE/κ	Legs	Lower arms	Neck	Trunk	Upper arms	Wrists	Mean
MLP-Skel2D	.157/.740	.056/.791	.291/.560	.238/.676	.169/.790	.149/.713	.177/.712
MLP-Skel3D	.079/.860	.041/.850	.174/.733	.148/.799	.128/.839	.129/.753	.117/.806
RN50	.129/.783	.072/.719	.268/.594	.250/.665	.201/.757	.175/.662	.183/.697
DN121	.123/.797	.059/.782	.226/.661	.221/.701	.167/.796	.161/.695	.160/.739
Ours-1	.121/.803	.060/.775	.239/.639	.234/.688	.165/.797	.149/.712	.161/.736
Ours-2	.119/.804	.057/.782	.241/.639	.236/.683	.160/.802	.150/.712	.161/.737
Ours-4	.112/.816	.056/.788	.240/.643	.234/.686	.150/.815	.155/.702	.158/.742
Ours-6	.112/.815	.056/.788	.245/.633	.234/.687	.152/.813	.153/.704	.159/.740

De todos los métodos basados en imágenes utilizados, el prospecto obtuvo los mejores resultados en todas las partes del cuerpo, con una mejora promedio del 5%. También se concluyó que entrenar a las seis partes de cuerpo juntas da mejores resultados que entrenar cada parte individualmente.

Debilidades

El tamaño de la red y la complejidad del entrenamiento y prueba aumentan con el número de ramas. Por lo que es importante seguir mejorando el modelo para reducir la complejidad y hacerlo más eficiente.

Ding, Z., Li, W., Jie Chi Yang, Ogunbona, P., & Qin, L. (2024). An attention-based CNN for automatic whole-body postural assessment. Expert Systems with Applications, 238, 122391–122391. https://doi.org/10.1016/j.eswa.2023.122391

Artículo 3 - Stress recognition from facial images in children during physiotherapy with serious games

Şilan Fidan Vural, Bengi Yurdusever, Ayse Betul Oktay, Ismail Uzun

Este estudio propone un método basado en aprendizaje profundo y de máquina para la detección de estrés en niños. Las contribuciones incluyen el uso de un conjunto de datos novedoso que comprende videos de 25 niños, incluidos niños con condiciones específicas, y la exploración del uso de imágenes faciales de adultos para la ampliación de datos. El método propuesto emplea una arquitectura de red neuronal profunda modificada, VGG-Face, para el reconocimiento de emociones faciales y utiliza tres modelos de aprendizaje automático para el reconocimiento del estrés. Este estudio representa el primer intento de reconocer el estrés en imágenes faciales de niños durante la fisioterapia con juegos. Los hallazgos tienen el potencial de optimizar los resultados del paciente y contribuir al monitoreo de los pacientes durante la terapia en el hogar.

A continuación se describen las etapas que nos son relevantes para el reto:

Dataset

A lo largo de este estudio, utilizamos el conjunto de datos AKTIVES para el reconocimiento del estrés y otros tres conjuntos de datos abiertos para el reconocimiento de emociones faciales. El resumen de los conjuntos de datos utilizados en este estudio se presenta en la Tabla 1.

Table 1. List of datasets used in this study.

Dataset	Sample	Subject	Ages	Labels
AKTIVES	50 videos	25	10.2 ±1.27	stressed or not
NIMH-Chefs	534 images	59	mean 13.6	fearful, angry, happy, sad and neutral
LIRIS-CSE	208 videos	12	6-12 years	Disgusted, happy, sad, surprised, neutral, feared
FER-2013	35,887 images	unknown	All ages	Disgusted, happy, sad, surprised, neutral, feared, angry

Metodología

En primera instancia se utilizó la red neuronal para la clasificación de emociones del dataset. Posteriormente se realizó el reconocimiento del estrés, este es una clasificación binaria. Se utiliza la salida del vector de emoción facial y las características de cada niño para la detección del estrés. Se emplean tres modelos de aprendizaje automático diferentes para la detección del estrés: K-NN (Cover & Hart, 1967), Árboles de Decisión (Wu et al., 2008) y XGBoost (Chen & Guestrin, 2016).

Las entradas de los modelos de aprendizaje automático son la enfermedad de los niños (normalmente plexo braquial, discapacidad intelectual y dislexia), el sexo (masculino o femenino) del niño, el ID del niño, el número de fotograma del video y la salida del modelo VGG-Face, que es un vector de 5 dimensiones que muestra las probabilidades de cada emoción.

Cada método de aprendizaje automático se prueba en el conjunto de datos AKTIVES, que tiene datos desbalanceados. Hay un total de 213,690 fotogramas, donde 191,037 de ellos están etiquetados como "no estresados" y 22,653 como "estresados". Se utiliza el 80% de los datos (173,088 fotogramas) para el entrenamiento y el 20% de los datos (42,738 fotogramas) para la prueba.

Métricas de evaluación del modelo

Para medir el rendimiento de los tres modelos se utiliza la matriz de confusión se muestra en la Figura 8. La precisión, la sensibilidad, la recuperación y los puntajes F1 de cada método se presentan en la Tabla 3.

Los datos de AKTIVES están desbalanceados y la matriz de confusión muestra que la clase de "no estresados" se reconoce con mayor precisión con cada método. El puntaje F1 de la clase "estresados" es del 70% con KNN, 80% con árboles de decisión y 81% con XGBoost. La precisión de la clase "estresados" con XGBoost es del 86% y la precisión de los Árboles de Decisión es del 78%. Sin embargo, la sensibilidad de la clase "estresados" con XGBoost es del 77% y la precisión de los Árboles de Decisión es del 83%. KNN tiene un puntaje F1 del 70% para la clase "estresados", siendo el más bajo de todos los métodos.

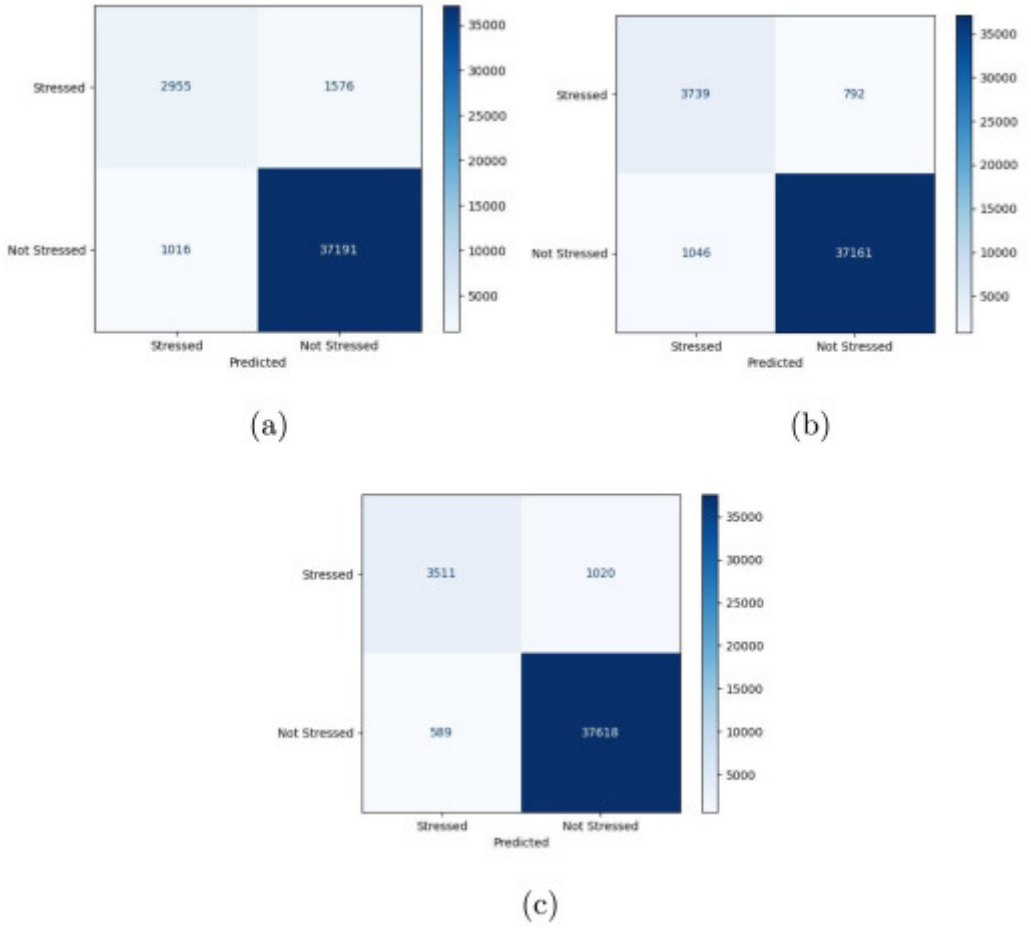


Fig. Matriz de Confusión del Reconocimiento de Estrés con (a) KNN, (b) Árboles de Decisión y (c) XGBoost.

KNN				
	Precision	Recall	F1 score	Support
Stressed	0.74	0.65	0.70	4531
Not Stressed	0.96	0.97	0.97	38207
Macro Avg	0.85	0.81	0.83	42738
Weighted Avg	0.94	0.94	0.94	42738
Decision Trees				
	Precision	Recall	F1 score	Support
Stressed	0.78	0.83	0.80	4531
Not Stressed	0.98	0.97	0.98	38207
Macro Avg	0.88	0.90	0.89	42738
Weighted Avg	0.96	0.96	0.96	42738
XGBoost				
	Precision	Recall	F1 score	Support
Stressed	0.86	0.77	0.81	4531
Not Stressed	0.97	0.98	0.98	38207
Macro Avg	0.91	0.88	0.90	42738
Weighted Avg	0.96	0.96	0.96	42738

Tabla 3. Resultados de la detección de estrés.

Este caso es un caso de clasificación en donde utilizan métricas de Precisión, Recall, F1 score y support para validar los resultados. Estas métricas nos son de utilidad para la clasificación, en nuestro caso tenemos una clasificación de estimación de pose para saber si una persona está participando o no.

Artículo 4 - Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model

El estudio "Enhancing Crop Productivity and Sustainability Through Disease Identification in Maize Leaves: Exploiting a Large Dataset with an Advanced Vision Transformer Model" introduce un modelo avanzado de transformador de visión, el MaxViT, adaptado y mejorado para la detección de enfermedades en hojas de maíz. El modelo, optimizado con un bloque Squeeze-and-Excitation y una MLP basada en Normalización de Respuesta Global, logra una precisión del 99.24% y una alta velocidad de inferencia. La investigación incluye una comparativa exhaustiva de más de 28 modelos CNN y 36 de transformadores de visión, utilizando el dataset más extenso creado a partir de PlantVillage, PlantDoc y CD&S. Este modelo avanzado es efectivo para aplicaciones prácticas en agricultura, demostrando un enfoque significativo en la identificación precisa y rápida de enfermedades en plantas.

Dataset

El CD&S dataset es un conjunto de datos de alta resolución que clasifica las enfermedades de las hojas de maíz en tres categorías: Mancha Gris de la Hoja, Roya del Norte y Mancha de la Hoja del Norte. Este conjunto no incluye hojas de maíz sanas y contiene un número específico de imágenes para cada categoría. Para evaluar la capacidad de generalización del modelo, se asignó aleatoriamente el 70% de los datos para entrenamiento, 15% para validación y 15% para pruebas. El nuevo dataset se construyó combinando clases idénticas de los datasets PlantVillage, PlantDoc y CD&S. Se incluyeron todas las clases de las enfermedades de PlantVillage y PlantDoc, y solo Mancha Gris de la Hoja y Roya del Norte del CD&S, excluyendo la clase de Mancha de la Hoja del Norte para evitar un desbalance en el dataset.

Metodología y Métricas

La metodología propuesta en el estudio se centra en la identificación oportuna y precisa de enfermedades en hojas de plantas mediante algoritmos de clasificación y detección de objetos basados en aprendizaje profundo. El modelo MaxViT, adaptado para identificar enfermedades en hojas de maíz, se ha escalado eficazmente para este propósito, utilizando un enfoque de atención múltiple y una estructura de bloques que incluye atención local y global, proporcionando una solución más eficiente con menos parámetros para un nuevo dataset de maíz de 4 clases.

Las arquitecturas de aprendizaje profundo han revolucionado la inteligencia artificial, especialmente en visión por computadora. Los CNNs son fundamentales en este campo, pero tienen limitaciones como su incapacidad para capturar información global en una imagen. Para superar esto, se han desarrollado los transformadores de visión, que utilizan mecanismos de autoatención para capturar dependencias de largo alcance en los datos.

El MaxViT es un modelo de aprendizaje profundo innovador en el campo de la visión por computadora. Su capacidad para manejar grandes conjuntos de datos visuales complejos y adaptarse a características específicas de entrada lo convierte en una herramienta poderosa para una amplia gama de tareas de visión por computadora. Su diseño secuencial y escalabilidad mejoran aún más su rendimiento y capacidades, posicionándolo como un fuerte contendiente en el campo de la visión por computadora.

Así mismo, a continuación se mostraran algunos valores de las métricas utilizadas en este reporte.

Details of some MaxViT models with Proposed MaxViT model.

Stage	Size	MaxViT-Proposed	MaxViT-Tiny	MaxViT-Small
S0: Stem Block	1/2	B = 2C = 64	B = 2C = 64	B = 2C = 64
S1: MaxViT-Block	1/4	B = 1C = 64	B = 2C = 64	B = 2C = 96
S2: MaxViT-Block	1/8	B = 2C = 128	B = 2C = 128	B = 2C = 192
S3: MaxViT-Block	1/16	B = 4C = 256	B = 5C = 256	B = 5C = 384
S4: MaxViT-Block	1/32	B = 1C = 512	B = 2C = 512	B = 2C = 768
Number of Parameters (M)		**19.07**	30.41	68.17

Tabla 4. Resultados de MaxVit.

La Tabla 4 muestra las configuraciones de diferentes variantes del modelo MaxViT para el procesamiento de imágenes en tareas de visión por computadora, específicamente para la identificación de enfermedades en hojas de maíz. La variante "MaxViT-Proposed" tiene significativamente menos parámetros (19.07 millones) en comparación con "MaxViT-Tiny" y "MaxViT-Small", que tienen 30.41 y 68.17 millones de parámetros, respectivamente. Esto sugiere que el modelo propuesto puede ser más eficiente computacionalmente, potencialmente ofreciendo una velocidad de inferencia más rápida y siendo más fácil de implementar en sistemas con recursos limitados, mientras mantiene un alto rendimiento en la tarea específica.

Results of CNN models on new maize (merged) dataset.

Model	Accuracy	Precision	Recall	F1-score
VGG-13 (Simonyan & Zisserman, 2015)	0.9733	0.9729	0.9736	0.9732
VGG-16	0.9733	0.9725	0.9751	0.9737
VGG-19	0.9733	0.9746	0.9728	0.9737
ResNet-18 (K. He et al., 2016)	0.9771	0.9776	0.9768	0.9772
ResNet-34	0.9758	0.9761	0.9758	0.9760
ResNet-50	0.9746	0.9749	0.9745	0.9747
ResNet-101	0.9733	0.9732	0.9741	0.9737
DenseNet-121 (Huang et al., 2016)	**0.9809**	**0.9810**	**0.9817**	**0.9813**
DenseNet-169	0.9733	0.9740	0.9731	0.9736
DenseNet-201	0.9720	0.9705	0.9749	0.9724
EfficientNetv2-small (Tan & Le, 2021)	0.9796	0.9794	0.9796	0.9795
EfficientNetv2-medium	0.9758	0.9773	0.9748	0.9760
EfficientNetv2-large	0.9783	0.9788	0.9781	0.9784
Xception (Chollet, 2016)	0.9784	0.9779	0.9786	0.9782
Inceptionv4 (Yu et al., 2023)	0.9682	0.9668	0.9724	0.9687
MobileNetv3-small (Howard et al., 2019)	0.9784	0.9792	0.9789	0.9790
MobileNetv3-large	0.9733	0.9750	0.9718	0.9732
RegNetx-120 (Xu et al., 2021)	0.9771	0.9785	0.9768	0.9776
RegNety-120	0.9771	0.9769	0.9779	0.9774
HrNetw18v2-Small (Wang et al., 2019)	0.9733	0.9737	0.9731	0.9734
PnasNet5-Large (C. Liu et al., 2017)	0.9746	0.9752	0.9736	0.9743
Res2Net50-26w-4 s (Gao et al., 2019)	0.9720	0.9720	0.9723	0.9721
Res2Net101-26w-4 s	0.9733	0.9744	0.9731	0.9737
Res2Net50-26w-6 s	0.9758	0.9777	0.9748	0.9761
Res2Net50-26w-8 s	0.9746	0.9741	0.9756	0.9749
Res2Net50-14w-8 s	0.9758	0.9761	0.9764	0.9762
Res2Next50	0.9771	0.9776	0.9771	0.9774
Res2Net101D	0.9771	0.9766	0.9782	0.9774

Tabla 5. Resultados de modelos CNN.

Finalmente, en la tabla 5 podemos observar el análisis de las métricas de rendimiento de varios modelos CNN en un dataset unificado de maíz muestra que la mayoría de los modelos tienen una alta exactitud (accuracy), precisión, recuperación (recall) y puntuación F1, con

valores que oscilan principalmente alrededor del 97%. El modelo EfficientNetV2-small destaca por su alta precisión y recuperación, lo que sugiere un equilibrio entre la capacidad de detectar positivos verdaderos y la precisión de dichas detecciones. El modelo ResNet-101 parece ofrecer el mejor rendimiento general basado en la puntuación F1. Estos resultados indican que los modelos CNN más avanzados y profundos tienden a tener un mejor rendimiento en el conjunto de datos de maíz fusionado, lo que puede ser una consideración importante para la selección del modelo en aplicaciones prácticas como la de nuestro reto.

De manera análoga a la aplicación de Convolutional Neural Networks (CNN) en la detección de enfermedades en hojas de maíz, la utilización de algoritmos avanzados, como dlib_face_recognition_resnet_model_v1, en el ámbito de Face Recognition en nuestro proyecto integrador representa una convergencia hacia enfoques más sofisticados en la identificación facial. Así como las CNN se destacan en la captura de patrones específicos en imágenes de hojas de maíz, el modelo ResNet con 29 capas convolucionales se especializa en la localización precisa de rostros y el emparejamiento con identidades específicas. Este enfoque es crucial dada la complejidad inherente a la detección facial en entornos no controlados, donde factores como variaciones en poses, expresiones faciales, iluminación y oclusión facial presentan desafíos similares a los encontrados en la identificación de enfermedades en plantas. La elección de algoritmos avanzados en ambas aplicaciones refleja la necesidad de abordar tareas complejas mediante enfoques que puedan aprender y adaptarse a patrones específicos en conjuntos de datos diversos.

Referencias:

Pacal, I. (2024). Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model. *Expert Systems With Applications*, 238, 122099. https://www.elsevier.com/locate/eswa

DeepMatcher:

Tao Xie, Kun Dai, Ke Wang, Ruifeng Li, Lijun Zhao,

DeepMatcher: A deep transformer-based network for robust and accurate local feature matching,

Expert Systems with Applications,

Volume 237, Part A,

2024,

121361,

ISSN 0957-4174,

https://doi.org/10.1016/j.eswa.2023.121361.

(https://www.sciencedirect.com/science/article/pii/S0957417423018638)

Abstract: Local feature matching constitutes the cornerstone of multiple computer vision applications (e.g., 3D reconstruction and long-term visual localization), and has been successfully resolved by detector-free methods. To further improve the matching performance, more recent research has focused on designing sophisticated architectures but endures additional computational overhead. In this study, with a different perspective from previous studies, we aim to develop a deep and compact matching network to improve performance while reducing computing cost. The key insight is that a local feature matcher with deep layers can capture more human-intuitive and simpler-to-match features. To this end, we propose DeepMatcher, a deep transformer-based network that tackles the inherent obstacles of not being able to build a deep local feature matcher with current methods.

DeepMatcher consists of: (1) a local feature extractor (LFE), (2) a feature-transition module (FTM), (3) a slimming transformer (SlimFormer), (4) a coarse matches module (CMM), and (5) a fine matches module (FMM). The LFE is utilized to generate dense keypoints with enriched features from the images. We then introduce the FTM to ensure a smooth transition of feature scopes from LFE to the subsequent SlimFormer because of their different receptive fields. Subsequently, we develop SlimFormer dedicated to DeepMatcher, which leverages vector-based attention to model the relevance among all keypoints, enabling the network to construct a deep Transformer architecture with less computational cost. Relative position encoding is applied to each SlimFormer to explicitly disclose the relative distance information, thereby improving the representation of the keypoints. A layer-scale strategy is also employed in each SlimFormer to enable the network to adaptively assimilate message exchange, thus endowing it to simulate human behavior, in which humans can acquire different matching cues each time they scan an image pair. By interleaving the self- and cross-SlimFormers multiple times, DeepMatcher can easily establish pixel-wise dense matches at the coarse level using the CMM. Finally, we consider match refinement as a combination of classification and regression problems and design an FMM to predict confidence and offset concurrently, thus generating robust and accurate matches. Compared with our baseline LoFTR in indoor/outdoor pose estimation, DeepMatcher surpasses it by 3.32%/2.91% in AUC@5◦. Besides, DeepMatcher and DeepMatcher-L significantly reduce computational cost and only consume 77.89% and 92.46% GFLOPs of LoFTR. Large DeepMatcher considerably outperforms state-of-the-art methods on several benchmarks, including outdoor pose estimation (MegaDepth dataset), indoor pose estimation (ScanNet dataset), homography estimation (HPatches dataset), and image matching (HPatches dataset), demonstrating the superior matching capability of a deep local feature matcher.

YOLO POSE BLOG:

https://deci.ai/blog/pose-estimation-yolo-nas-pose/

Keywords: Deep learning; Local feature matching; Efficient transformer

Vural, S., Yurdusever, B., Oktay, A. B., & Uzun, I. (2024). Stress recognition from facial images in children during physiotherapy with serious games. *Expert Systems with Applications*, *238*, 121837. https://doi.org/10.1016/j.eswa.2023.121837