Universidad de Lima

Facultad de Ingeniería

Carrera de Ingeniería de Sistemas

COMPARATIVA ENTRE RESNET-50, VGG-16, VISION TRANSFORMER Y SWIN TRANSFORMER PARA EL RECONOCIMIENTO FACIAL CON OCLUSIÓN DE UNA MASCARILLA

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Brenda Xiomara Tafur Acenjo

Código 20172692

Martin Alexis Tello Pariona

Código 20163654

Asesor

Edwin Jonathan Escobedo Cardenas

Lima – Perú

Noviembre de 2023

Comparativa entre RESNET-50, VGG-16, Vision Transformer y Swin Transformer para el reconocimiento facial con oclusión de una mascarilla

Brenda Xiomara Tafur Acenjo
20172692@aloe.ulima.edu.pe
Universidad de Lima

Martin Alexis Tello Pariona
20163654@aloe.ulima.edu.pe
Universidad de Lima

Resumen: En el contexto de la pandemia, el reconocimiento facial tomó importancia al ser un método de verificación sin contacto físico. Es así que esta investigación evaluó el *accuracy* de modelos preentrenados (VGG-16, RESNET-50, Vision Transformer, Swin Transformer) para la verificación de identidad, enfrentando el desafío de la oclusión por mascarilla. Los resultados revelaron que los modelos *Transformers* superaron a las CNN en *accuracy*. Este trabajo contribuye significativamente al explorar dos tipos de arquitecturas y al crear un conjunto de datos público, enriqueciendo la investigación en visión computacional para el reconocimiento facial con oclusión de mascarilla.

Palabras Clave: Reconocimiento facial, Vision Transformer, Swin transformer

Abstract: In the context of the pandemic, face recognition became important as a non-physical contact verification method. Thus, this research evaluated the accuracy of pre-trained models (VGG-16, RESNET-50, Vision Transformer, Swin Transformer) for identity verification, facing the challenge of mask occlusion. The results revealed that the transformer models outperformed the CNNs in accuracy. This work contributes significantly by exploring two types of architectures and creating a public dataset, enriching computer vision research for face recognition with mask occlusion.

Keywords: Face Recognition, Vision Transformer, Swin transformer

Línea de investigación IDIC – ULIMA

- Calidad de vida y bienestar: Salud

Área y Sub-áreas de Investigación:

- Computing Methodologies
- Computer Vision

Objetivo (s) de Desarrollo Sostenible (ODS)

- ODS 3: Salud y bienestar
- ODS 9: Industria, innovación e infraestructura
- ODS 11: Ciudades y comunidades sostenibles

1. PLANTEAMIENTO DEL PROBLEMA

La pandemia propició la demanda por soluciones de verificación biométrica sin contacto físico. Sin embargo, el desafío fue que se hiciera el reconocimiento facial ante un rostro parcialmente ocluido. Los tipos de oclusión más estudiados eran aquellos generados por el uso de gafas de sol, bufandas, cabello en el rostro, envejecimiento, etc. (Sáez Trigueros et al., 2018). Además, otra limitante para ese entonces era las pocas bases de datos de dominio público de personas con y sin mascarilla. A pesar de que existían conjuntos de datos libres de rostros enmascarados como Real World Fasked Face Recognition Dataset (RMFRD, por sus siglas en inglés), estos solo eran de acceso para un grupo reducido de la industria y la academia. Así, el público en general no podía usarlos ni investigarlos sin restricciones (Laxminarayanamma et al., 2021).

Debido a lo mencionado, se propuso realizar una comparativa en el accuracy de cuatro modelos preentrenados de reconocimiento facial: VGG-16 y Resnet-50, Vision Transformer y Swin Transformer. Todo ello con la finalidad de cuantificar la métrica señalada en distintos escenarios y experimentos. Para esto se creó un conjunto de datos propio.

2. OBJETIVO

Determinar el accuracy de los modelos CNN (VGG-16 y RESNET-50), y los modelos Transformers (ViT y Swin), ante la oclusión de una mascarilla facial. Se busca comparar el rendimiento de estas arquitecturas en escenarios sin mascarilla y con mascarilla, con el fin de identificar cuál ofrece mejores resultados en la verificación de identidad en situaciones de obstrucción. Asimismo, la creación de una base de datos con el volumen necesario de imágenes para el entrenamiento de estos modelos.

3. JUSTIFICACIÓN

La necesidad de verificar la identidad sin contacto físico se volvió esencial en el contexto de la pandemia. Una solución a ello fue el reconocimiento facial; sin embargo, la oclusión parcial del rostro representó un desafío a superar. Asimismo, crear un conjunto de datos de dominio público se vio necesario debido a la escasez de bases de datos con mascarilla en ese momento. Contar con una base de datos pública fomenta la replicabilidad de los experimentos, impulsando el avance en el campo de la visión computacional. Además, los resultados obtenidos ofrecen información clave para tomar decisiones informadas sobre qué modelos son más apropiados para implementar sistemas de verificación de identidad sin contacto en situaciones de oclusión.

La investigación se centra en identificar al mejor modelo de reconocimiento facial en espacios cerrados usando una mascarilla. Esto se relaciona con el ODS 3, que busca garantizar una vida saludable y promover el bienestar para todos en todas las edades. De igual manera, el trabajo se enfoca en evaluar modelos preentrenados de reconocimiento facial en un contexto de uso innovador para abordar los desafíos creados por la pandemia. Esto se alinea con el ODS 9, que promueve la construcción de infraestructuras resilientes, la promoción de la industrialización inclusiva y sostenible, y el fomento de la innovación. Por último, se aborda la necesidad de ideas y soluciones innovadoras para espacios cerrados, lo que se relaciona con el ODS 11, que busca hacer las ciudades y los asentamientos humanos inclusivos, seguros, resilientes y sostenibles.

4. DISEÑO METODOLÓGICO

Como primer paso se definieron las directrices para las capturas de videos. Una vez concluida la recolección de videos con nuestros voluntarios, se procedió con la fase de preprocesamiento de imágenes en la cual se desglosaron los videos en fotogramas. Con el objetivo de mejorar la robustez de nuestros modelos, se consideraron variaciones para los gestos que realizan los colaboradores en las grabaciones, esto nos permitió capturar una mayor cantidad de características propias del colaborador en diferentes ángulos, lo que resulta fundamental para el reconocimiento facial (Damer et al. 2020).

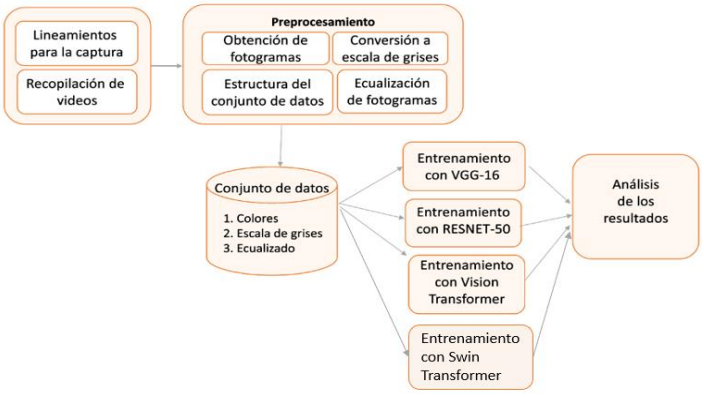
Para la fase de preprocesamiento, se empleó la red convolucional MTCNN para procesar las imágenes. A continuación, se diseñó la estructura del conjunto de datos y se extrajeron fotogramas de los vídeos. Se obtuvieron alrededor de 1200 imágenes por vídeo por cada persona, como se detalla en el trabajo de Yanai & Kawano (2015). Como resultado de la etapa de preprocesamiento se obtuvo un conjunto de datos con imágenes a colores, a escala de grises y ecualizadas.

Luego, se realizaron los entrenamientos con los modelos escogidos: VGG-16 y RESNET-50, Vision Transformer y Swin Transformer. Cabe destacar que se opta por usar la transferencia de aprendizaje, ya que estos modelos son preentrenados sobre conjuntos de datos enormes, lo cual hace que el modelo tenga pesos ya establecidos para clasificar una imagen y con alto grado de accuracy. Otro punto determinante, es el ahorro en recursos computacionales y tiempo que demandaría entrenar desde cero cada modelo que se ha presentado en el artículo. Para llevar a cabo los entrenamientos, se utilizó la plataforma Google Colab, la cual contiene las siguientes especificaciones técnicas con las que se entrenan los modelos: Procesador de modelo Intel(R) Xeon(R) CPU @ 2.20GHz, 1 Core, 6ta Generación, 2200.186 MHz, 12.7 GB de RAM, 225 GB de HDD, Sistema Operativo Linux. Asimismo, se utiliza una GPU de modelo NVIDIA-SMI, Versión del Driver 525.85.12 y 12va Versión CUDA.

Finalmente con los resultados obtenidos se realizó la comparación entre los modelos bajo la métrica de accuracy. La figura 1 muestra cada una de las etapas del proceso.

Figura 1

Etapas de la metodología



AGRADECIMIENTOS

A nuestro asesor Edwin Escobedo, padres y comunidad que participó en la elaboración de esta tesis.

REFERENCIAS

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Ge, Y., Liu, H., Du, J., Li, Z., & Wei, Y. (2023). Masked face recognition with convolutional visual self-attention network. Neurocomputing, 518, 496-506.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Kumar, G. (2023). Smartphone Authentication with Lightweight Deep Learning.

Datos del artículo publicado

- Nombre del artículo: Comparativa entre RESNET-50, VGG-16, Vision Transformer y Swin Transformer para el reconocimiento facial con oclusión de una mascarilla
- Autores: Brenda Tafur y Martin Tello
- Co autor(es): Edwin Escobedo

Publicación en revista

- Nombre de la revista: Interfases
- Volumen: 1
- Número: 17
- Año: 2023
- Pp: 56 - 78
- Enlace web donde se encuentra publicado el artículo (identificador DOI, ISBN, ISSN o equivalentes): https://revistas.ulima.edu.pe/index.php/Interfases/article/view/6361

entrega 2

INFORME DE ORIGINALIDAD

17%	16%	9%	11%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	www.scholink.org Fuente de Internet	2%
2	Yujia Xu, Hak-Keung Lam, Guangyu Jia. "MANet: A two-stage deep learning method for classification of COVID-19 from Chest X-ray images", Neurocomputing, 2021 Publicación	2%
3	Submitted to Universidad EAN Trabajo del estudiante	2%
4	repositorioacademico.upc.edu.pe Fuente de Internet	2%
5	colaboracion.dnp.gov.co Fuente de Internet	1%
6	hdl.handle.net Fuente de Internet	1%
7	repository.javeriana.edu.co Fuente de Internet	1%
8	ijournalse.org Fuente de Internet	1%