Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Momento de Retroalimentación: Reto Evaluación

TC3007C.501 Inteligencia Artificial Avanzada para la Ciencia de Datos II

Profesores:

Iván Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernández

Félix Ricardo Botello Urrutia

Edgar Covantes Osuna

Felipe Castillo Rendón

Hugo Terashima Marín

Equipo 2

Integrantes:

Luis Ángel Guzmán Iribe - A01741757

Julian Lawrence Gil Soares - A00832272

Alberto H Orozco Ramos - A00831719

8 de Noviembre de 2023

1. **Realicen una búsqueda de literatura reciente acerca de las métricas utilizadas por otros autores para evaluar el desempeño de un modelo, en problemas similares (puede ser la identificación de rostro, reconocimiento de participación, estimación de pose, etc.). Se recomienda que realicen la búsqueda a través de Scopus (https://www.scopus.com/home.uriLinks to an external site.)**
2. **Seleccionen por lo menos cuatro (4) artículos y realicen un resumen (un párrafo corto por artículo) donde mencionan lo que los autores hicieron, la métrica que utilizaron, y alguna debilidad de la solución propuesta por los autores**

Durante nuestra investigación encontramos que para los modelos de reconocimiento de objetos se utiliza la precisión como la principal métrica de evaluación. Específicamente utilizamos una comparación de los verdaderos positivos y verdaderos negativos contra los falsos positivos y falsos negativos. En nuestra investigación encontramos la siguiente ecuación:

$$Precisión = \frac{TP}{TP-FP} .$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1\ score = \frac{2*recall*precisión}{recall+precision}$$

Estas ecuaciones nos permiten evaluar el desempeño de nuestro modelo a través de medir cuántas veces nuestro modelo atina en su reconocimiento.

Para explicar algunos de los conceptos y fórmulas mencionadas, tenemos las siguientes explicaciones para que resulte más comprensible la interpretación de dichas métricas:

- **Matriz de Confusión (*Confusion Matrix*)**: La matriz de confusión es una tabla que provee una descripción detallada del rendimiento del modelo al proyectar las cuentas de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Además, permite identificar errores específicos que el modelo presenta. Este recurso resulta ser muy valioso, especialmente cuando se necesita evaluar y mejorar el rendimiento del modelo.

$$TP = Verdaderos\ positivos$$
$$FP = Falsos\ positivos$$
$$TN = Verdaderos\ negativos$$
$$FN = Falsos\ negativos$$

- **Exactitud (*Accuracy*)**: Este mide la proporción de instancias que fueron clasificadas correctamente del resto por el modelo. Básicamente, proporciona una medida general de cómo es el modelo, en términos de clasificar correctamente tanto verdaderos positivos y verdaderos negativos.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precisión (*Precision*):** Mide la precisión del modelo con respecto a predicciones positivas. Da respuesta a la pregunta "De todas las instancias estimadas por el modelo ¿cuántas de ellas fueron positivas?" Entre mayor sea este valor, indica que el modelo es bueno al evitar falsos positivos.

$$Precision = \frac{TP}{TP + FP}$$

- **Exhaustividad (*Recall*)**: Este mide la habilidad del modelo para identificar todas las instancias relevantes dentro del dataset. Un valor de "recall" alto indica que el modelo raramente se equivoca en casos positivos. Esta métrica es importante cuando los casos positivos faltantes son bastante costosos, un ejemplo serían los diagnósticos médicos.

$$Recall = \frac{TP}{TP + FN}$$

- **Valor F1 (*F1-Score*)**: La métrica F1 es la media armónica de la precisión con respecto a la exhaustividad. Esto condensa el intercambio de precisión y recall en una sola métrica. Un alto valor de "F1-Score" indica que el modelo demuestra un rendimiento destacado en términos de las variables que considera (precision y recall) Resulta muy útil cuando se busca balancear la precisión y la exhaustividad o bien para identificar una distribución de métricas desigual.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Otra metodología de evaluación de desempeño que encontramos que se utiliza como un estándar es el llamado Intersection Over Union (IOU). La manera en la que se utiliza se puede describir con la siguiente ecuación: $IOU = \frac{Area\ of\ Union}{Area\ of\ Intersection}$. Este método evalúa la ubicación y tamaño de la caja que la computadora pone encima del objeto que se detectó. El área de unión se refiere al área total del área de el objeto real dentro de la imagen y la caja de la precisión que el algoritmo arroja, el área de intercesión se refiere a cuánta área del objeto real está dentro de la caja de previsión. Su división forma un rango de 0 a 1 que representa básicamente que tan certero es la predicción arrojada y este número se puede utilizar para evaluar el rendimiento del algoritmo.

En un artículo (De Lima Medeiros et al., 2022) que trata acerca del uso de computer vision y modelos de deep learning para la detección de parpadeos voluntarios en pacientes que padecen de esclerosis lateral amiotrófica. El artículo trata con la implementación de un modelo que funciona en diferentes capas o fases, de manera similar al nuestro, con la primera fase siendo la detección del ojo, la cual emplea la técnica de IOU descrita anteriormente para el entrenamiento del modelo detección de posición del ojo. Posteriormente, se realizan diferentes pruebas de video para evaluar la capacidad del modelo, en las que se define un parpadeo como la superposición del IOU en el sobre la muestra en el dataset, es decir, cuando se detecta una modificación en el área de interés en una determinada cantidad de tiempo. Una vez contabilizados los parpadeos se buscan las superposiciones en el tiempo con la muestra real, y se obtiene la matriz de confusión que da

a conocer los resultados positivos y negativos, verdaderos y falsos, que a su vez permiten conocer las métricas de desempeño descritas anteriormente.

Otro artículo (Goh et al., 2023) que se revisó habla sobre el desarrollo de aplicaciones web que implementan deep learning, su respectivo deployment, enfoque de desarrollo y al final se mencionan algunos ejemplos. Resumiendo un poco de lo que trata el artículo, en primera instancia se abordan las aplicaciones web que implementan modelos de Deep Learning del lado del cliente en los navegadores. El autor comparte su opinión diciendo que este tipo aplicaciones poseen muchas ventajas como amplia búsqueda, decremento en costos para servidores y privacidad de los datos de los usuarios. También, hay secciones del documento dónde se discute el porqué los navegadores, el lenguaje de programación JavaScript y la biblioteca de TensorFlow.js son ideales para este conjunto, así como las restricciones y consideraciones involucradas. Igualmente, se describen 4 enfoques para la creación y optimización de modelos de Deep Learning que mejor se ajusten para aplicaciones web por el lado del frontend. Dichos enfoques son: *reusar*, *personalizar*, *convertir* y *optimizar* modelos existentes, y resaltar sus ventajas y desventajas. Finalmente, en el artículo contiene una sección donde incluye algunos proyectos de desarrollo web que hacen uso de Deep Learning, tales como *“Education playground Teachable Machine”*, *“Visualization playground GAN Lab”* o *“Development playground Milo”*, sin embargo, existen algunas en particular que poseen elementos que utilizamos en el reto tales como detección de pose y gestos. Algunos ejemplos que se mencionan son *“Rehabilitation and monitoring”*, *“Physical activity coaching”*, y *“Gesture tracking”*. Dichos ejemplos utilizan principalmente el modelo de pose.
“PoseNet” para estimar la postura y los gestos de las personas en diferentes escenarios, ya sea para rehabilitación, monitoreo, evaluación de riesgos, detección de caídas, baile y entrenamiento físico; dichos modelos hacen uso de dos cámaras o la integración de voz para mejorar la interacción y precisión de los resultados.

En cuánto a la relación entre los modelos propuestos por el autor y nuestro modelo existen diversos puntos a considerar. Por ejemplo, nosotros si hacemos uso de servicios en la nube para implementar los modelos de Deep Learning pre entrenados que utilizamos para la toma de asistencia y participación de los alumnos. Otra diferencia es que el autor menciona el uso de Tensorflow.js para la simplificación durante la implementación de los modelos, cosa que no contemplamos desde un inicio con el fin de evitar recurrir a servidores de la nube para ejecutar los modelos y elevar el costo computacional. Algunas de las métricas que utilizan este tipo de modelos son:

- *Focal Loss Function:* Función de pérdida computarizada para manejar la pérdida de datos por cada ejemplo analizado y reajusta los pesos del modelo acorde a su dificultad, la cual se ajusta al uso de su entrenamiento.
- *Mean Average Precision y IoU:* Estas métricas utilizadas para evaluar la exactitud del modelo miden el porcentaje de las predicciones correctas por cada una de los valores a predecir que son desconocidos por el modelo. Por otro lado, IoU evalúa la superposición entre los valores predecidos y las cajas delimitadoras de verdad.

3. **Seleccionen métricas (o indicadores de desempeño) a medir adecuadas al reto y justifiquen su selección.**
   Ya que nuestro modelo se centra principalmente en la asignación de participaciones a alumnos a través del reconocimiento de poses, decidimos someter a prueba la capacidad

del modelo de identificar las participaciones dentro de una imagen, y asignar dicha participación a la persona adecuada.

Para realizar la evaluación del modelo, tomamos en cuenta los siguientes escenarios para la realización de la matriz de confusión:

- Verdadero Positivo: la persona X levantó la mano, y se detectó y asignó la participación de manera correcta.
- Verdadero Negativo: la persona X no levantó la mano, y se no se detectó ninguna participación.
- Falso Positivo: la persona X no levantó la mano, sin embargo, se detectó y asignó una participación a dicha persona.
- Falso Negativo: la persona X levantó la mano, sin embargo, no se detectó su participación, o se asignó a una persona equivocada.

O visto de otro modo:

	Persona X levantó la mano	Persona X no levantó la mano
Modelo predijo que la persona X levantó la mano	VP	FP
Modelo predijo que la persona X no levantó la mano	FN	VN

Consideramos que para este escenario en particular, es necesario evaluar la capacidad del modelo todos los casos positivos posibles, en esta situación, no registrar una participación tiene un mayor peso que registrar una participación inexistente, es decir, los falsos negativos tiene un peso mayor que los falsos positivos.

De acuerdo con el sitio web Evidentlyai.com (2023), la métrica de recall es la más apropiada para evaluar modelos en esta clase de situaciones, ya que le otorga un mayor peso a la capacidad del sistema de identificar todos los casos positivos, dentro de los casos de prueba. Recapitulando, el recall se calcula de la siguiente forma:

$$Recall = \frac{TP}{TP + FN}$$

En términos conceptuales, esto representa la proporción de casos positivos que fue capaz de detectar el modelo, que resulta ser precisamente lo que estamos buscando. Realizaremos este proceso para cada una de las personas, y luego sumaremos los resultados individuales de cada para evaluar el comportamiento del modelo en general.

4. **Evalúen los modelos usando las métricas (o indicadores de desempeño).**
   En este caso empleamos 5 videos de ejemplo, de los cuales extraemos fotogramas que consideramos de interés para probar diferentes escenarios. Se buscó probar diferentes posiciones y agregar dificultad adicional para el reconocimiento agregando elementos como

gafas, cabello que cubre el rostro, o capuchas para dificultar el reconocimiento, y acercarnos más a un escenario de un salón real.

Los videos de los que se obtienen dichos fotogramas se encuentran en esta carpeta de Google Drive: Modelo Outputs

5. **Generen comparaciones entre los modelos con las métricas (o indicadores de desempeño).**
   Debido a la falta de tiempo y recursos, desarrollamos una sola versión del modelo, por lo que no podemos tomar un punto de comparación con otros modelos; sin embargo, consideramos que este ejercicio es una buena actividad para conocer las áreas de oportunidad del modelo, y conocer los puntos en los que es necesario centrar esfuerzos para mejorarlo en futuras iteraciones del proyecto.

6. **Grafiquen y visualicen los resultados obtenidos.**
   Estos fueron los resultados que arrojó el modelo para cada una de las personas involucradas.

   Luis Ángel:

	Luis Ángel levantó la mano	Luis Ángel no levantó la mano
Modelo predijo que Luis Ángel levantó la mano	3	2
Modelo predijo que Luis Ángel no levantó la mano	1	2

   Alberto:

	Alberto levantó la mano	Alberto no levantó la mano
Modelo predijo que Alberto levantó la mano	2	3
Modelo predijo que Alberto no levantó la mano	1	5

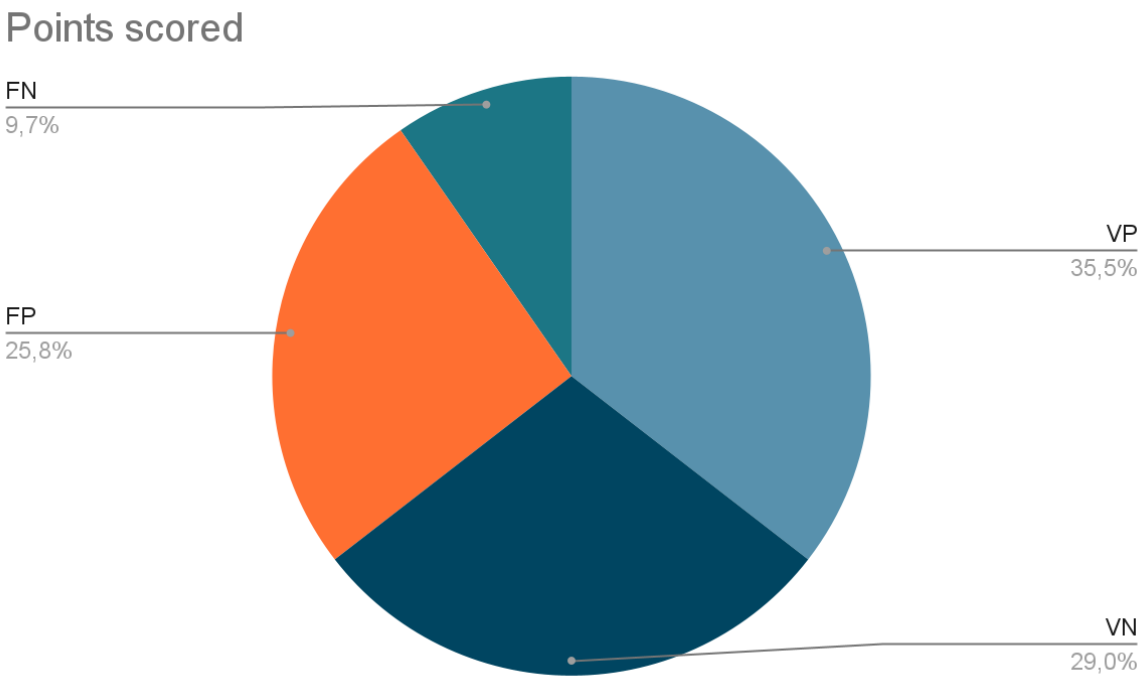
   Antonio:

	Antonio levantó la mano	Antonio no levantó la mano
Modelo predijo que Antonio levantó la mano	3	3
Modelo predijo que Antonio	1	2

no levantó la mano		

Total:

	Persona X levantó la mano	Persona X no levantó la mano
Modelo predijo que la persona X levantó la mano	11	8
Modelo predijo que la persona X no levantó la mano	3	9



Tomando estos datos en cuenta, obtenemos el siguiente Recall:

$$Recall = \frac{TP}{TP + FN} = \frac{11}{11 + 3} = 0.78$$

En adición, calculamos las métricas adicionales para tener una idea más completa del desempeño del modelo:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{11 + 8}{11 + 9 + 8 + 3} = 0.61$$

$$Precision = \frac{TP}{TP + FP} = \frac{11}{11 + 9} = 0.55$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot 0.55 \cdot 0.78}{0.55 + 0.78} = \frac{0.858}{1.33} = 0.64$$

7. **Utilicen los elementos anteriores para poder hacer una interpretación clara y de calidad**

   En base a los resultados obtenidos con las métricas de evaluación del modelo, podemos afirmar que este tiende a sobreestimar la toma de participaciones, esto lo podemos observar en casos como por ejemplo, cuando un estudiante levanta el brazo para que se le considere la participación, el modelo en algunas ocasiones tiende a registrar más de una participación, aún y cuando existe una restricción que debería de impedir este tipo de comportamientos. Por otro lado, el modelo suele marcar participaciones a las personas equivocadas cuando no debería existir algún problema para determinar quién es la persona que levantó el brazo.

   Este último hecho puede indicar que el modelo de reconocimiento falla en algunas circunstancias o bien, el modelo de pose falla de alguna forma en casos específicos. Sea cual sea la razón, estos resultados nos permitirán realizar los debidos ajustes al modelo de forma que sea posible reducir el margen de error y aumentar estas métricas al menos en un 10% cada una para iteraciones posteriores.

   También es importante rescatar que aunque es posible mejorar el Recall, este muestra que el modelo es capaz de reconocer la mayoría de instancias positivas del conjunto de datos de prueba, que es precisamente la estadística que buscamos maximizar, ya que consideramos que tiene un peso más elevado el hecho de no reconocer una participación, a otorgar participaciones adicionales.

   En general, si bien el modelo cuenta con áreas de mejora, estas están identificadas, y nos permitirían implementar medidas para evitar que los fallos surjan con tanta frecuencia, y garantizar que no se pierdan los resultados positivos.

.

Referencias

- Farkhod, A., Abdusalomov, A. B., Mukhiddinov, M., & Cho, Y.-I. (2022). Development of Real-Time Landmark-Based Emotion Recognition CNN for Masked Faces. Sensors (14248220), 22(22), 8704. https://0-doi-org.biblioteca-ils.tec.mx/10.3390/s22228704
- Goh, H.-A., Ho, C.-K., & Abas, F. S. (2023). Front-end deep learning web apps development and deployment: a review. Applied Intelligence, 53(1), 15923-15945.
- Jeune, P. L., & Mokraoui, A. (2023). Rethinking Intersection Over Union for Small Object Detection in Few-Shot Regime.
- De Lima Medeiros, P. A., Da Silva, G. V. S., Fernandes, F., Sánchez-Gendriz, I., De Castro Lins, H. W., Da Silva Barros, D. M., Nagem, D. a. P., & De Medeiros Valentim, R. A. (2022). Efficient machine learning approach for volunteer eye-blink detection in real-time using webcam. *Expert Systems With Applications*, *188*, 116073. https://doi.org/10.1016/j.eswa.2021.116073
- Accuracy vs. precision vs. recall in machine learning: What's the difference? (2023). Evidentlyai. Recuperado 27 de noviembre de 2023, de https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall