

Video análisis aplicado al seguimiento de jugadores de fútbol

Alexander Leonardo Lique Lamas



alexander.lique.l@uni.pe
Departamento de Ciencia en Computación,
Universidad Nacional de Ingeniería
Asesor: Marcos Antonio Alania Vicente

26 de Enero, 2023



1 Introducción

2 Marco Conceptual

3 Metodología

4 Experimentos y Resultados

5 Conclusiones y Trabajos Futuros



Fútbol



Figura: Brazil vs Serbia, Qatar 2022¹

¹Frames extraídos de:

<https://www.youtube.com/watch?v=H-Z64PJpbBY>



Video Análisis



Figura: Video análisis en el Fútbol²

²The role of video analysis in football :

<https://www.youtube.com/watch?v=aCrJzG1ctas&t>



Problemática

- ▶ Para realizar un análisis de video en el fútbol es necesario hacer un seguimiento de los jugadores, lo que conlleva a capturar su posición en el campo.



Problemática

- ▶ Para realizar un análisis de video en el fútbol es necesario hacer un seguimiento de los jugadores, lo que conlleva a capturar su posición en el campo.
- ▶ La falta de algoritmos precisos de deep learning para la detección es una problemática en el campo de la inteligencia artificial ya que aún están en investigación.



Objetivos Generales

El objetivo de este seminario es presentar la arquitectura de la vision transformer DETR y aplicarla a la detección de jugadores en videos de fútbol.



Objetivos Específico

- Realizar un estudio sobre la arquitectura Detection Transformer y compararlo con Faster R-CNN.



Objetivos Específico

- ▶ Realizar un estudio sobre la arquitectura Detection Transformer y compararlo con Faster R-CNN.
- ▶ Entrenar dichos modelos sobre el dataset de la competición de SoccerNet.



Objetivos Específico

- ▶ Realizar un estudio sobre la arquitectura Detection Transformer y compararlo con Faster R-CNN.
- ▶ Entrenar dichos modelos sobre el dataset de la competición de SoccerNet.
- ▶ Comparar dichas arquitecturas según las métricas establecidas en COCO, que fueron usadas en el paper original de DETR.



Convolución

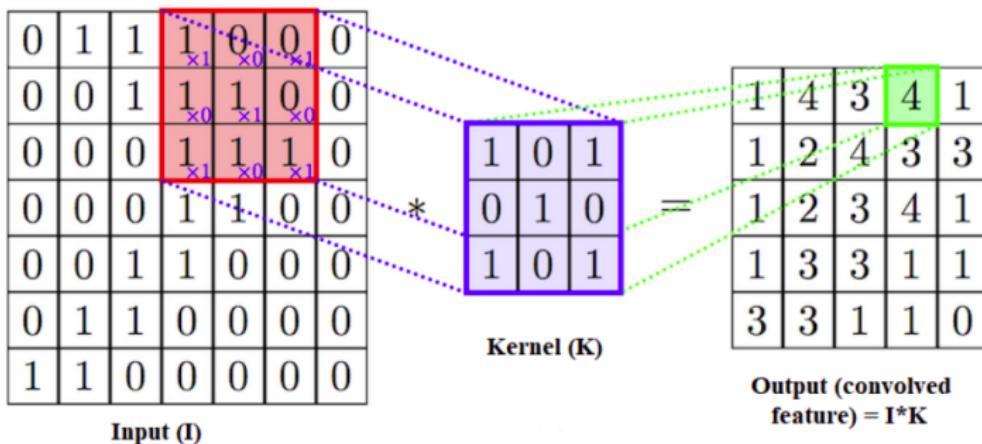


Figura: Operación de convolución (Sultana et al. 2019)



Convolución

Dependiendo del tipo de filtro que nosotros escojamos. Obtendremos diferentes tipos de efectos tales como detectar bordes, suavizado, eliminación de ruido, etc.

-1	-1	-1
-1	8	-1
-1	-1	-1

outline ▾



Figura: Filtro outline³

³Esta imagen fue creada gracias a
<https://setosa.io/ev/image-kernels/>



Residual Network (ResNet)

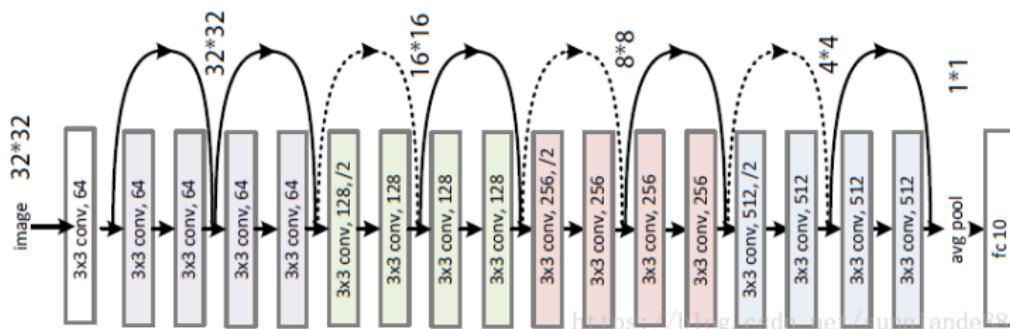


Figura: ResNet18 (Duchesnay et al. s.f.)



Feature Pyramid Networks

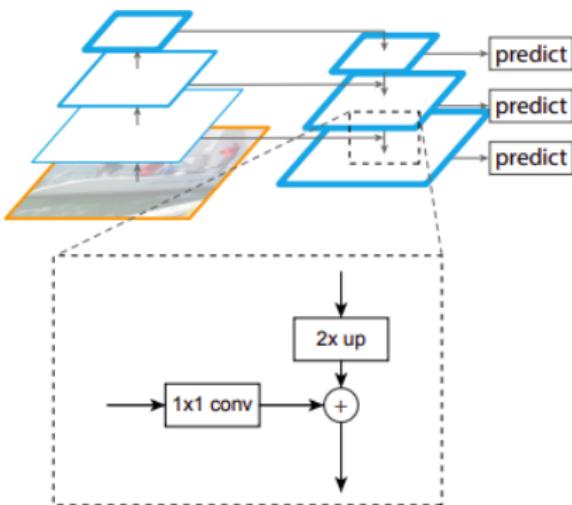


Figura: Arquitectura FPN (Lin et al. 2016)



ResNET + Feature Pyramid Networks

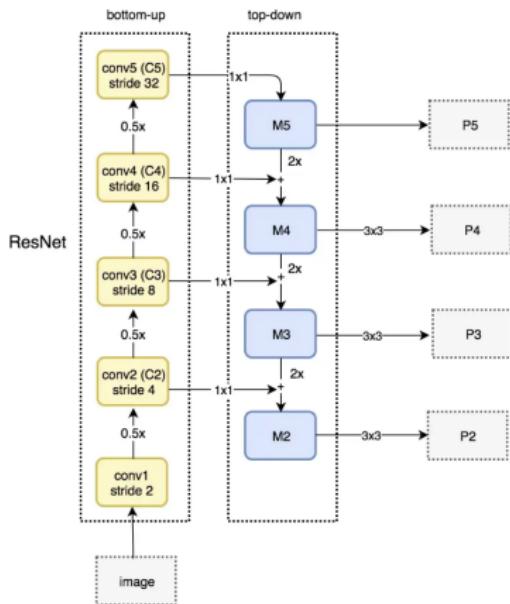


Figura: Arquitectura ResNet + FPN ⁴

⁴Understanding Feature Pyramid Networks for object detection (FPN) Jonathan Hui



Faster R-CNN

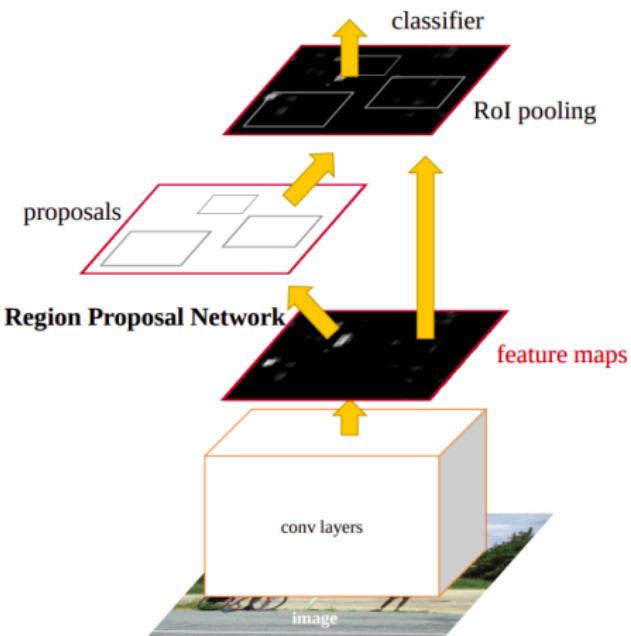


Figura: Faster R-CNN (Ren et al. 2015)



Codificación posicional

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (1)$$

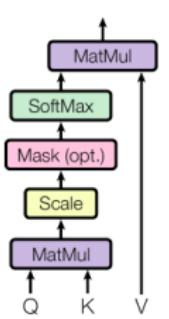
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (2)$$

- ▶ **pos** Posición de la palabra en la oración
- ▶ **i**: representa las diferentes dimensiones del codificador posicional y los embeddings.
- ▶ d_{model} : Tamaño del *word embedding*; como la del codificador posicional



Mecanismos de Atención

Scaled Dot-Product Attention



Multi-Head Attention

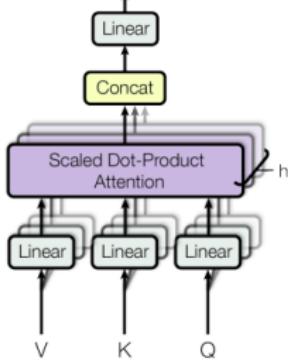


Figura: (izquierda) Atención de producto punto escalado. (derecha) *Multi-head attention* consta de varios capas de atención que se ejecutan en paralelo.(Vaswani et al. 2017)



Transformers

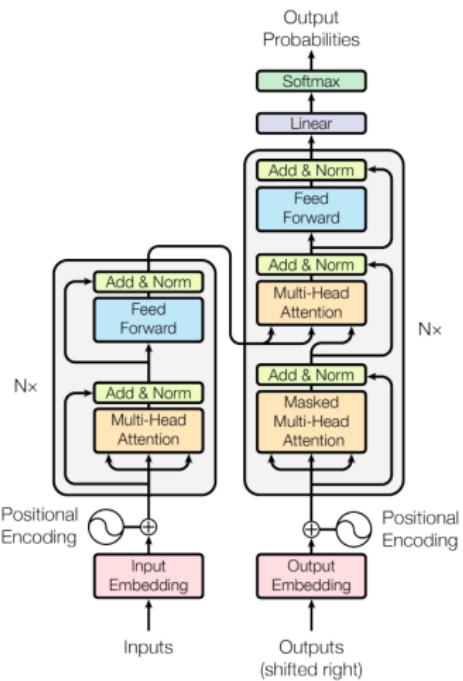


Figura: Transformers (Vaswani et al. 2017)



Detection Transformer(DETR)

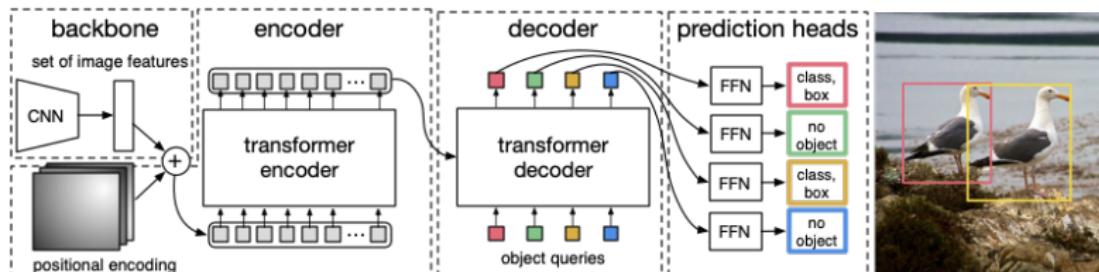


Figura: Detection Transformer (Carion et al. 2020)



Detection Transformer(DETR)

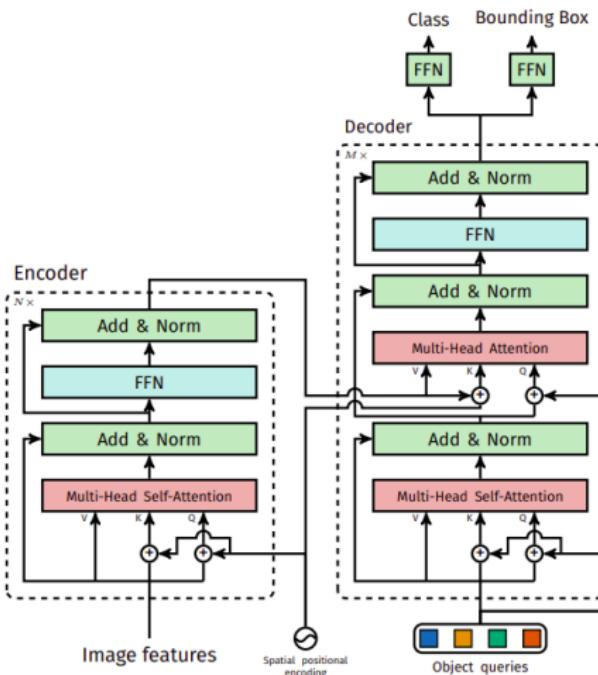


Figura: Arquitectura de DETR (Carion et al. 2020)

Detection Transformer

Mecanismo self-attention sobre los píxeles que conforman al objeto que está detectando el encoder.

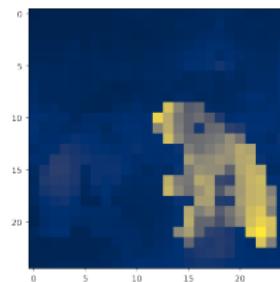
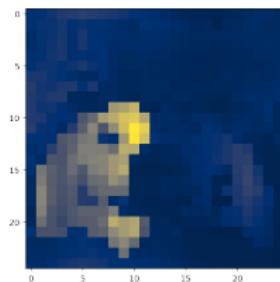


Figura: Visualización de self-attention del codificador⁵

⁵Esta imagen fue creada gracias al Google Colaboratory de Facebook research, *DETR's hands on Colab Notebook*. https://colab.research.google.com/github/facebookresearch/detr/blob/colab/notebooks/detr_attention.ipynb#V1OPEeVOYTEV



Metodología

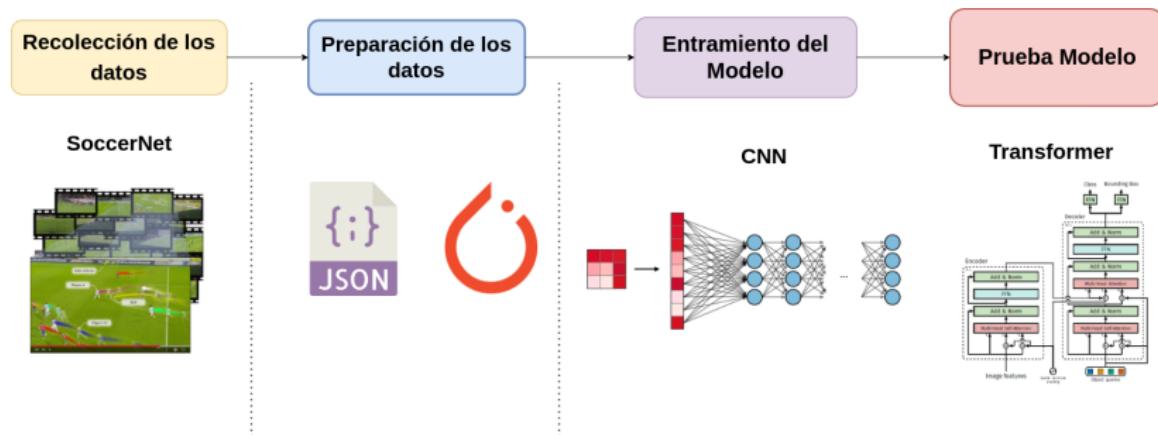


Figura: Pipeline del trabajo



SoccerNet



What is SoccerNet?

SoccerNet is a large-scale dataset for soccer video understanding. It has evolved over the years to include various tasks such as action spotting, camera calibration, player re-identification and tracking. It is composed of 550 complete broadcast soccer games and 12 single camera games taken from the major European leagues. SoccerNet is not only dataset, but also yearly challenges where the best teams compete at the International level. Join the SoccerNet community on [Discord](#) to discover how SoccerNet can further push your own research and industry.

News

- [2022-10] The SoccerNet team collaborated with [AinSports](#) and won the 2nd place of the Thakaa competition in AI+Sports!
- [2022-10] The results of the [SoccerNet 2022 Challenge](#) were presented at the [ACM Workshop on Multimedia Content Analysis in Sports](#).

Figura: Competición de SoccerNet⁶

⁶Link para descargar el conjunto de datos de SoccerNet:
<https://www.soccer-net.org>



Intersection Over Union (IoU)

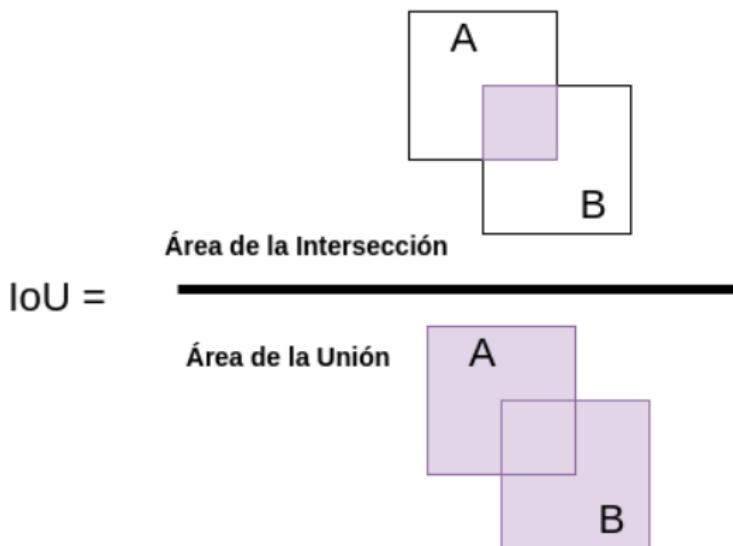


Figura: Intersección sobre la unión, imagen adaptada de (Géron s.f.)



Mean Average Pesción (mAP)

- ▶ Se calcula la matriz de confusión.

$u = 0.5$

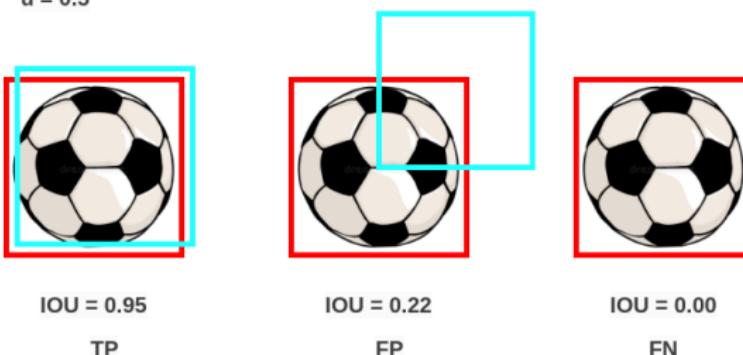


Figura: Análisis del IoU bajo un umbral $u = 0.5$



Mean Average Pescion (mAP)

- ▶ Se calcula la matriz de confusión.
- ▶ Calcular el Precision y Recall.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$



Mean Average Pescicion (mAP)

- ▶ Se calcula la matriz de confusión.
- ▶ Calcular el Precision y Recall.
- ▶ Se calcula el área bajo la curva de precisión-recall (AP)



Mean Average Pescion (mAP)

- ▶ Se calcula la matriz de confusión.
- ▶ Calcular el Precision y Recall.
- ▶ Se calcula el área bajo la curva de precisión-recall (AP)
- ▶ **Se calcula el AP sobre todas las clases detectadas.⁷**

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (5)$$



⁷Formula extraída de (Zhu et al. 2020)

Common Objects in Context (COCO)

Average Precision (AP):

AP % AP at IoU=.50:.05:.95 (primary challenge metric)
APIoU=.50 % AP at IoU=.50 (PASCAL VOC metric)
APIoU=.75 % AP at IoU=.75 (strict metric)

AP Across Scales:

APsmall % AP for small objects: area < 32^2
APmedium % AP for medium objects: 32^2 < area < 96^2
APlarge % AP for large objects: area > 96^2

Average Recall (AR):

AR^{max=1} % AR given 1 detection per image
AR^{max=10} % AR given 10 detections per image
AR^{max=100} % AR given 100 detections per image

AR Across Scales:

AR_{small} % AR for small objects: area < 32^2
AR_{medium} % AR for medium objects: 32^2 < area < 96^2
AR_{large} % AR for large objects: area > 96^2

Figura: Métricas de COCO⁸



⁸Métricas de la competición de COCO

<https://cocodataset.org/#detection-eval>

Herramientas

- ▶ Google colaboratory
- ▶ GPU Tesla T4 y CUDA v11.2.
- ▶ Se usó el Framework de Pytorch.



(a) Google colaboratory



(b) Pytorch



Experimentos

En el experimento se utilizaron 9073 imágenes para el entrenamiento, 1072 imágenes para la validación y 1288 para el test.

Modelo	#épocas	#lotes	Taza de aprendizaje	Optimizador
Faster RCNN+ResNet50+FPN	5	8	2e-5	Adam
DETR	20	8	1e-5	AdamW

Figura: Hiperparámetros y Optimizadores usados en el experimento



Tabla de Resultados

Model	#parámetros	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN+ResNet50+FPN	41076761	55.74	91.74	61.22	26.64	59.25	57.52
DETR	41286406	49.53	88.97	51.15	17.48	54.07	65.19

Figura: Resultados del entrenamiento de los modelos bajo la métrica de COCO



Resultados



(a) Faster R-CNN



(b) DETR

Figura: Inferencia de Faster R-CNN y DETR sobre un frame de SoccerNet test.



Resultados



(a) Faster R-CNN



(b) DETR

Figura: Inferencia de Faster RCNN y DETR sobre Perú vs Croacia ⁹



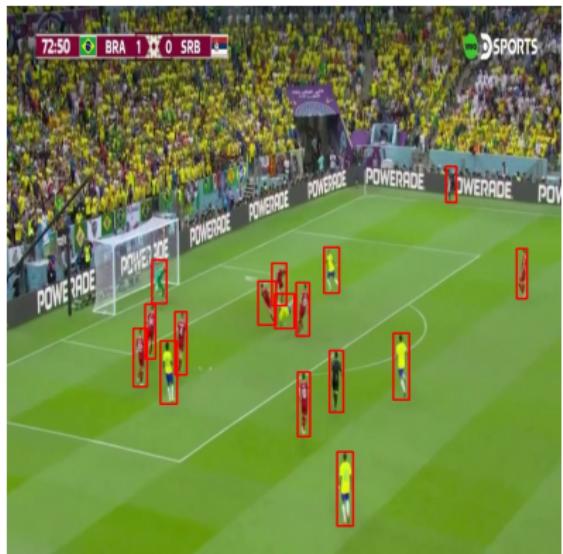
⁹Frames extraídos de:

<https://www.youtube.com/watch?v=O6ahXzs5j08>

Video análisis aplicado al seguimiento de jugadores de fútbol

Alexander Leonardo Lique Lamas

Resultados



(a) Faster R-CNN



(b) DETR

Figura: Brazil vs Serbia, Mundial Qatar 2022¹⁰



¹⁰Frames extraídos de

<https://www.youtube.com/watch?v=azsAKz5GBnO>

Resultados Faster



(a) Frame 1



(b) Frame 2

Figura: Inferencia del Argentina vs Croacia, Mundial Qatar 2022 ¹¹



¹¹Frames extraídos de:

<https://www.youtube.com/watch?v=Y1FBKUNAxwY>

Resultados Faster



(a) Frame 3



(b) Frame 4

Figura: Inferencia del Argentina vs Croacia, Mundial Qatar 2022 ¹²



¹²Frames extraídos de:

<https://www.youtube.com/watch?v=Y1FBKUNAxwY>

Resultados DETR



(a) Frame 1



(b) Frame 2

Figura: Inferencia del Argentina vs Croacia, Mundial Qatar 2022 ¹³



¹³Frames extraídos de:

<https://www.youtube.com/watch?v=Y1FBKUNAxwY>

Resultados DETR



(a) Frame 3



(b) Frame 4

Figura: Inferencia del Argentina vs Croacia, Mundial Qatar 2022 ¹⁴



¹⁴Frames extraídos de:

<https://www.youtube.com/watch?v=Y1FBKUNAxwY>

Conclusión

- ▶ Se muestra el buen desempeño de la arquitectura Faster R-CNN en comparación con DETR sobre el dataset de SoccerNet debido a su mayor estabilidad en la convergencia y su adaptabilidad a este conjunto de datos.



Conclusión

- ▶ Se muestra el buen desempeño de la arquitectura Faster R-CNN en comparación con DETR sobre el dataset de SoccerNet debido a su mayor estabilidad en la convergencia y su adaptabilidad a este conjunto de datos.
- ▶ **Se muestra el potencial que poseen los mecanismos de atención de los vision transformers para la detección.**



Trabajos Futuros

Basándose en la investigación realizada en este seminario de tesis y de cara a futuras investigaciones, propongo las siguientes opciones:

- Entrenar los modelos para realizar una detección más especializada para que distinga entre dos equipos o netamente del balón.



Trabajos Futuros

Basándose en la investigación realizada en este seminario de tesis y de cara a futuras investigaciones, propongo las siguientes opciones:

- ▶ Entrenar los modelos para realizar una detección más especializada para que distinga entre dos equipos o netamente del balón.
- ▶ Seguir ampliando el estudio sobre modelos más avanzados de *Vision transformer* para la detección .



Trabajos Futuros

Basándose en la investigación realizada en este seminario de tesis y de cara a futuras investigaciones, propongo las siguientes opciones:

- ▶ Entrenar los modelos para realizar una detección más especializada para que distinga entre dos equipos o netamente del balón.
- ▶ Seguir ampliando el estudio sobre modelos más avanzados de *Vision transformer* para la detección .
- ▶ También se planea reducir el tiempo de inferencia de estos modelos para el *Real-time* investigando técnicas de destilación.



Trabajos Futuros

Basándose en la investigación realizada en este seminario de tesis y de cara a futuras investigaciones, propongo las siguientes opciones:

- ▶ Entrenar los modelos para realizar una detección más especializada para que distinga entre dos equipos o netamente del balón.
- ▶ Seguir ampliando el estudio sobre modelos más avanzados de *Vision transformer* para la detección .
- ▶ También se planea reducir el tiempo de inferencia de estos modelos para el *Real-time* investigando técnicas de destilación.
- ▶ **A futuro se plantea investigar algoritmos para el *trackeo* de jugadores que hagan uso de estos algoritmo de detección.**



- [1] F. Sultana, A. Sufian y P. Dutta, "Advancements in Image Classification using Convolutional Neural Network," *CoRR*, vol. abs/1905.03288, 2019. arXiv: 1905.03288. dirección: <http://arxiv.org/abs/1905.03288>.
- [2] E. Duchesnay, T. Löfstedt y F. Younes, *Convolutional neural network¶*, dirección:
https://duchesnay.github.io/pystatsml/deep_learning/dl_cnn_cifar10_pytorch.html.
- [3] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan y S. J. Belongie, "Feature Pyramid Networks for Object Detection," *CoRR*, vol. abs/1612.03144, 2016. arXiv: 1612.03144. dirección:
<http://arxiv.org/abs/1612.03144>.
- [4] S. Ren, K. He, R. B. Girshick y J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *CoRR*, vol. abs/1506.01497, 2015. arXiv: 1506.01497. dirección: <http://arxiv.org/abs/1506.01497>.



- [5] A. Vaswani et al., "Attention Is All You Need," *CoRR*, vol. abs/1706.03762, 2017. arXiv: 1706.03762. dirección: <http://arxiv.org/abs/1706.03762>.
- [6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov y S. Zagoruyko, "End-to-end object detection with transformers," en *European conference on computer vision*, Springer, 2020, págs. 213-229.
- [7] A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems O'Reilly Media, 2017. 574 p," URL: <https://dl.acm.org/citation.cfm>, pág. 485,
- [8] H. Zhu, H. Wei, B. Li, X. Yuan y N. Kehtarnavaz, "A Review of Video Object Detection: Datasets, Metrics and Methods," *Applied Sciences*, vol. 10, n.º 21, 2020, ISSN: 2076-3417. DOI: 10.3390/app10217834. dirección: <https://www.mdpi.com/2076-3417/10/21/7834>.

