Utilización de Knowledge Distillation para entrenar un modelo más preciso

Uso de IA en trampas inteligentes para monitoreo en tiempo real

Matteo Martinez, Est. de Ing. en Mecatrónica.

1. Introducción al Problema

- Se desarrolló una trampa inteligente que utiliza IA para contar mosquitos.

- Necesidad de monitoreo en tiempo real requiere modelos rápidos y eficientes.

2. Dataset

El conjunto de datos original consistió en 1 199 imágenes reales, obtenidas del IICS. Estas imágenes fueron anotadas para entrenar un modelo de detección de mosquitos.

Distribución del Dataset:

- 240 imágenes se reservaron para validación.
- 240 imágenes se destinaron a pruebas (test).
- Las 719 imágenes restantes se utilizaron para entrenamiento.

Augmentation

Para enriquecer la diversidad del conjunto de entrenamiento y mejorar la capacidad de generalización del modelo, se aplicaron técnicas de data augmentation para tener tres variantes de cada imagen original, lo que elevó el número total de imágenes de entrenamiento a 2 157.

Técnicas de Augmentation Aplicadas:

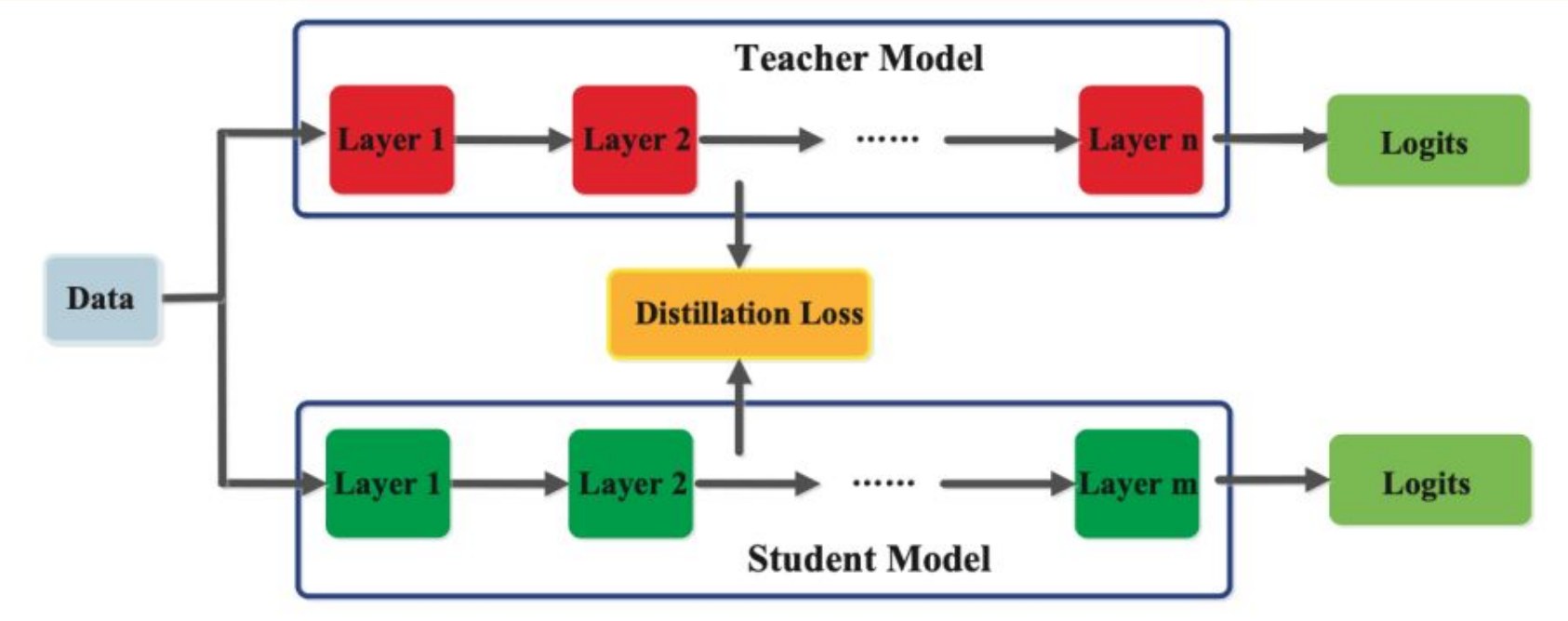
- Flips: Horizontal y vertical.
- Rotaciones 90°: Horaria, antihoraria y 180°.
- Rotaciones suaves: Entre -45° y +45°.
- Shear (corte): ±10° horizontal y vertical.
- Ajustes de saturación: Entre -25% y +25%.
- Brillo: Entre -15% y +15%.
- Exposición: Entre -10% y +10%.
- Desenfoque (blur): Hasta 1.2 px.
- Ruido: Hasta el 3.43% de los píxeles.

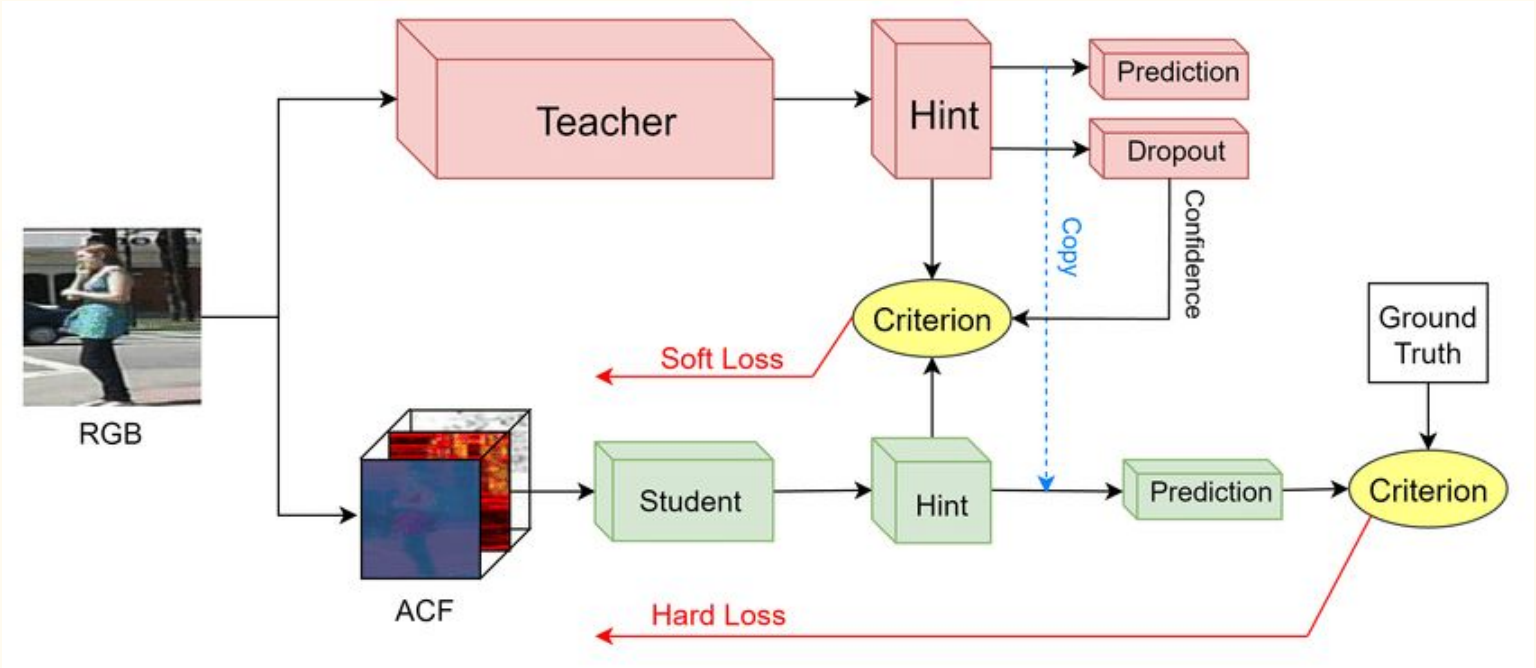
Balance de datos

Clases	Entrenamiento	Validacion	Confianza > 40%
All	2157	240	1861
Female	1159	129	1000
Incomplete	126	14	108
Male	916	102	790
Out of Focus	566	63	488
Undefined	161	18	138

3. Knowledge Distillation

- En el aprendizaje automático, la destilación o destilación del modelo del conocimiento es el proceso de transferencia de conocimientos de un modelo grande a uno más pequeño. Si bien los grandes modelos tienen más capacidad de conocimiento que los modelos pequeños, esta capacidad podría no utilizarse plenamente. Puede ser igual de costoso computacional evaluar un modelo incluso si utiliza poco de su capacidad de conocimiento. La destilación del conocimiento transfiere el conocimiento de un modelo grande a uno más pequeño sin pérdida de validez. Como los modelos más pequeños son menos costosos de evaluar, se pueden implementar en hardware menos potente.


Teacher Model
Layer 1
Layer 2
.....
Layer n
Logits
Data
Distillation Loss
Layer 1
Layer 2
.....
Layer m
Logits
Student Model


Teacher
Hint
Prediction
Dropout
Copy
Confidence
Criterion
Soft Loss
Ground Truth
RGB
ACF
Student
Hint
Prediction
Criterion
Hard Loss

Atributo	Valor aproximado
Parámetros	~3.2 millones
Tamaño del modelo	~6 MB
FPS (Velocidad)	+100 FPS en GPU moderna
Capas	~160 capas
Neuronas (aprox.)	1 a 1.2 millones de activaciones (neurona-locales)

Yolov8 Small como Teacher y Yolov8 Nano como Student

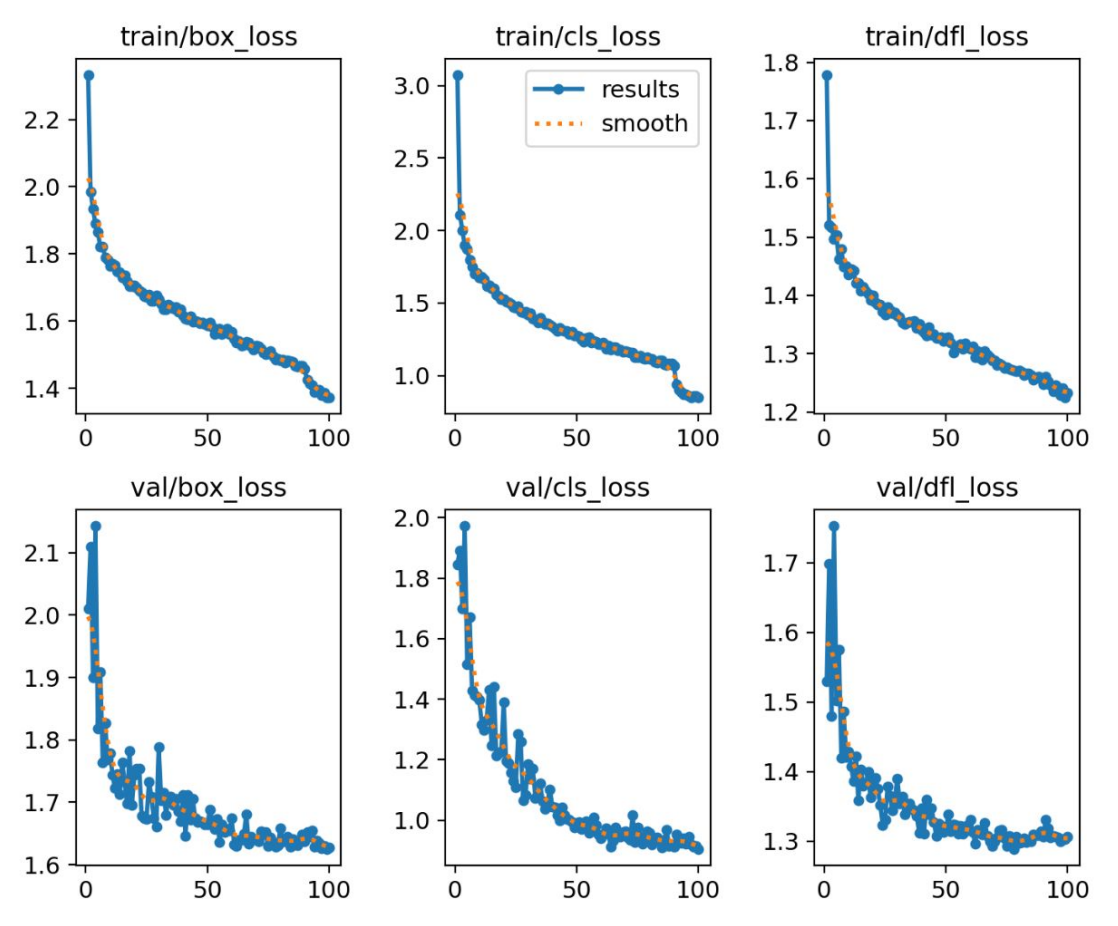
El objetivo de la implementación de la técnica Knowledge Distillation es disminuir el tiempo de inferencia del modelo, pues se requiere que este se acerque lo más posible al tiempo real, en especial en hardware movil.

Se realizaron 3 pruebas para comparar la efectividad de la técnica:

- Un modelo nano entrenado normalmente en 100 épocas.
- Un modelo nano entrenado con knowledge distillation en 100 épocas.
- Un modelo nano entrenado con knowledge distillation usando de student el modelo student ya entrenado anteriormente, y seleccionando en la base de datos las predicciones con un índice de confiabilidad superior a 40%.

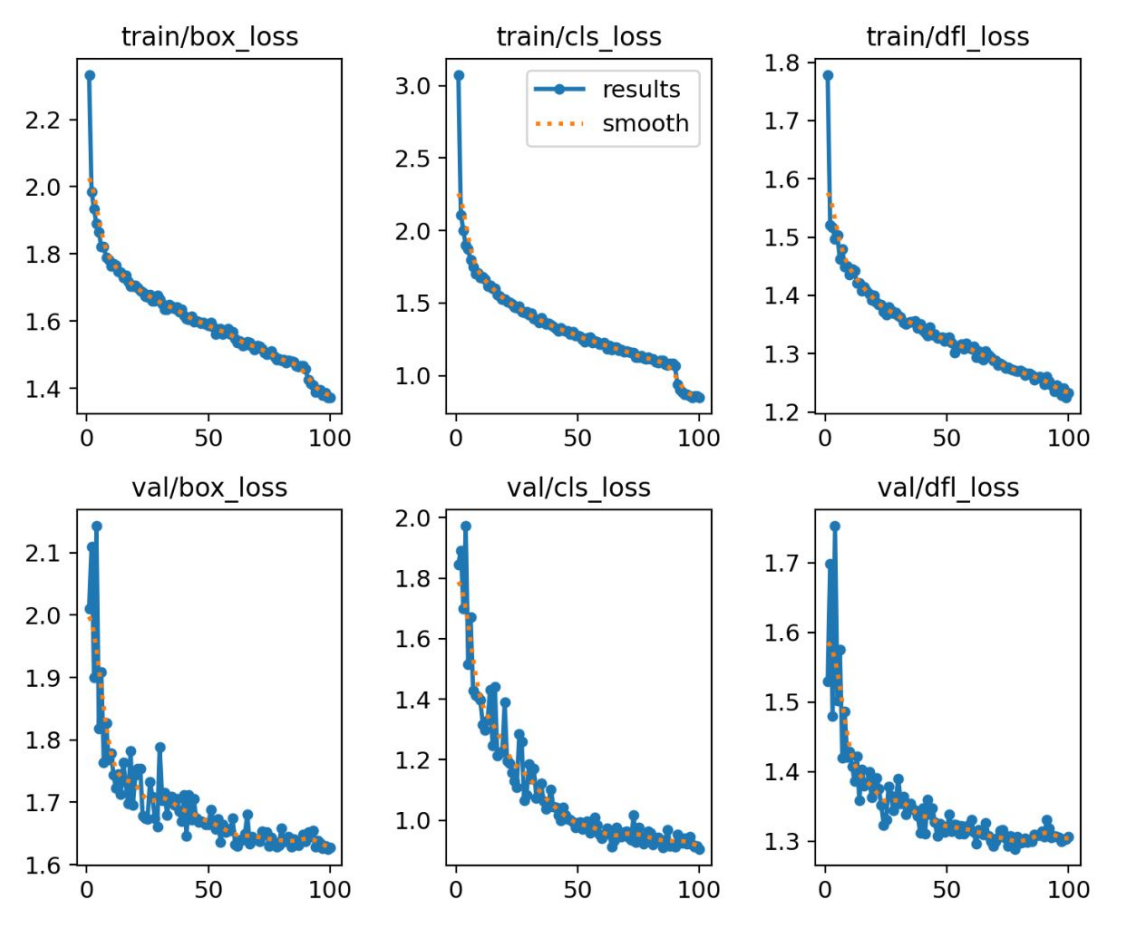
Parámetros de Entrenamiento	Normal	Knowledge Distillation	Knowledge Distillation re-entrenado
Img Size	224	224	224
Batch Size	16	16	16
Epochs	100	100	100
Learning Rate	0.01	0.01	0.01

Imagen 1: *Pérdidas del primer modelo entrenado con 100 épocas.*



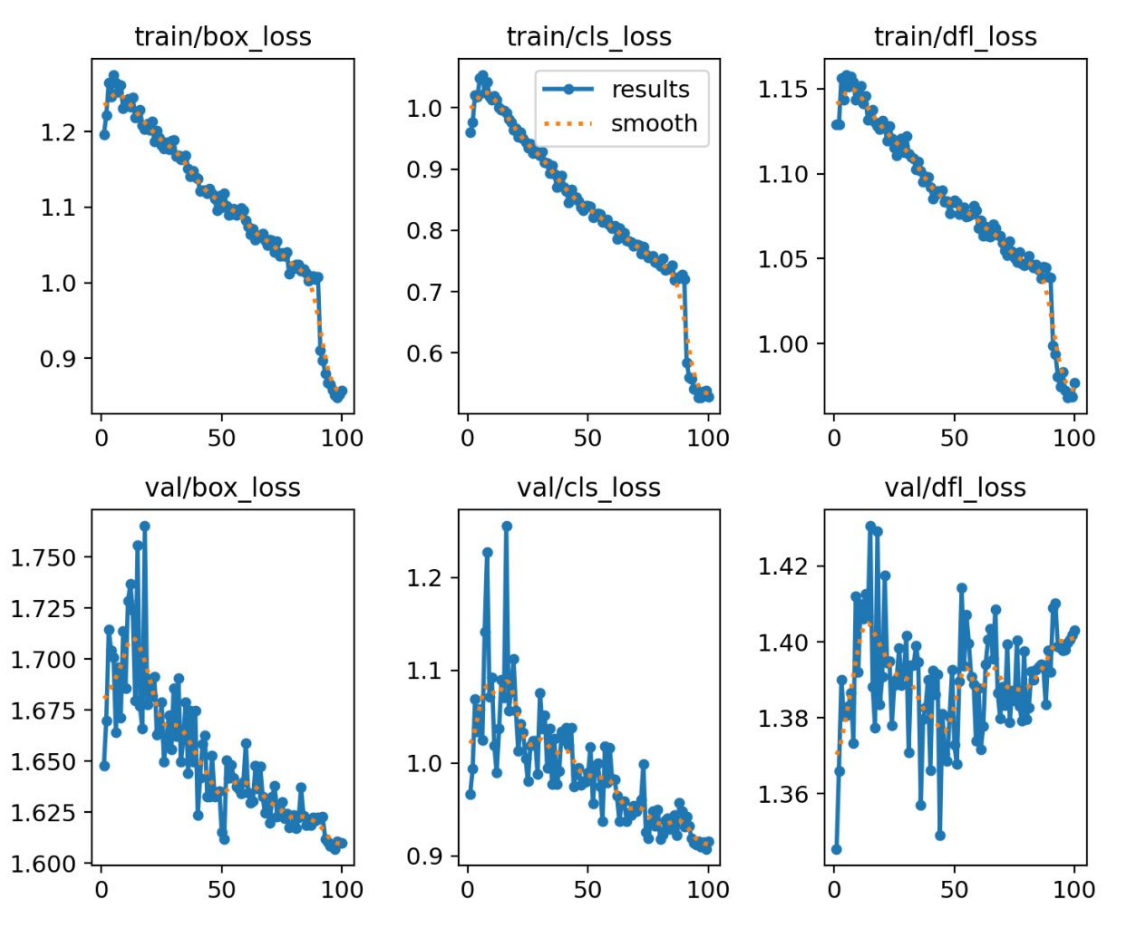
Fuente: datos propios.

Imagen 2: *Pérdidas del segundo modelo entrenado con 100 épocas.*



Fuente: datos propios.

Imagen 3: *Pérdidas del tercer modelo entrenado con 100 épocas.*



Fuente: datos propios.

4. Rendimiento Modelo Nano

- Se buscó el menor tiempo de inferencia y la mejor precisión.

Tabla 1: *mAP y tiempo de inferencia en para modelos nanos*

Modelo	mAP50 para Hembras	mAP50-95 para Hembras	mAP50 para Machos	mAP50-95 para Machos	Tiempo de Inferencia Promedio (ms)
Normal	87.1%	44.9%	93.9%	52.4%	4.5
Knowledge Distillation	87.1%	44.9%	93.9%	52.4%	4.4
Knowledge Distillation re-entrenado	88.7%	47.6%	94.1%	54.6%	4.6

Nota: Los tiempos corresponden a una GPU NVIDIA RTX 3090 con 24 GB de memoria.

Fuente: datos propios.

5. Conclusiones

- El Knowledge Distillation no influye el tiempo de inferencia
- El Knowledge Distillation puede mejorar la precisión del modelo a con un pequeño costo el el tiempo de inferencia.
- Próximo paso: despliegue en campo.

¡Muchas Gracias!