Optimización de Hiperparámetros OCR con Ray Tune para Documentos Académicos en Español

Sergio Jiménez Jiménez

Director: Javier Rodrigo Villazón Terrazas

Máster Universitario en Inteligencia Artificial

2025

unir LA UNIVERSIDAD EN INTERNET

ÍNDICE

Motivación

- La digitalización documental es una **necesidad estratégica** para organizaciones
- OCR como puente entre el mundo físico y digital
- Documentos en español: caracteres especiales ausentes en conjuntos de entrenamiento internacionales
- Modelos preentrenados: **rendimiento subóptimo** fuera de benchmarks estándar
- Fine-tuning requiere infraestructura costosa y datos etiquetados

Errores típicos en español

Original	OCR	Error
más	mas	Pérdida de acento
año	ano	Pérdida de eñe
¿Cómo	Como	Signos especiales
titulación	titulacióon	Duplicación

Planteamiento del Problema

¿Es posible mejorar significativamente el rendimiento de modelos OCR preentrenados para documentos en español mediante la optimización sistemática de hiperparámetros, sin requerir fine-tuning?

	Fine-tuning completo	Optimización de hiperparámetros
Datos	Miles de imágenes etiquetadas	Subconjunto de validación
Hardware	GPU alta memoria (>16 GB)	CPU / GPU consumo
Tiempo	Días / semanas	Minutos / horas
Expertise	Alto (ML avanzado)	Bajo-medio
Riesgo	Sobreajuste, catastrófico	Limitado, reversible

Objetivos

Objetivo general: Optimizar PaddleOCR para documentos académicos en español alcanzando un **CER < 2%** sin fine-tuning del modelo base.

- **OE1:** Comparar tres motores OCR open-source (EasyOCR, PaddleOCR, DocTR)
- **OE2:** Preparar dataset de evaluación de 45 páginas con ground truth
- **OE3:** Identificar hiperparámetros críticos mediante análisis de correlación
- **OE4:** Ejecutar 64 trials de optimización con Ray Tune + Optuna
- **OE5:** Validar la configuración optimizada frente al baseline

Estado del Arte: Motores OCR

EasyOCR

JaidedAI

CRAFT + CRNN

- 80+ idiomas
- Fácil de usar
- Baja configurabilidad

PaddleOCR

Baidu / PaddlePaddle

DB + SVTR (PP-OCRv5)

- Alta configurabilidad
- Pipeline modular
- Soporte español dedicado

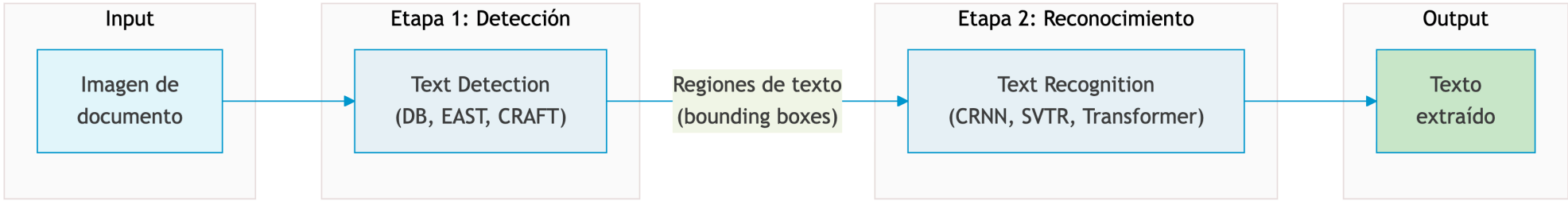
DocTR

Mindee

DB/LinkNet + CRNN/SAR

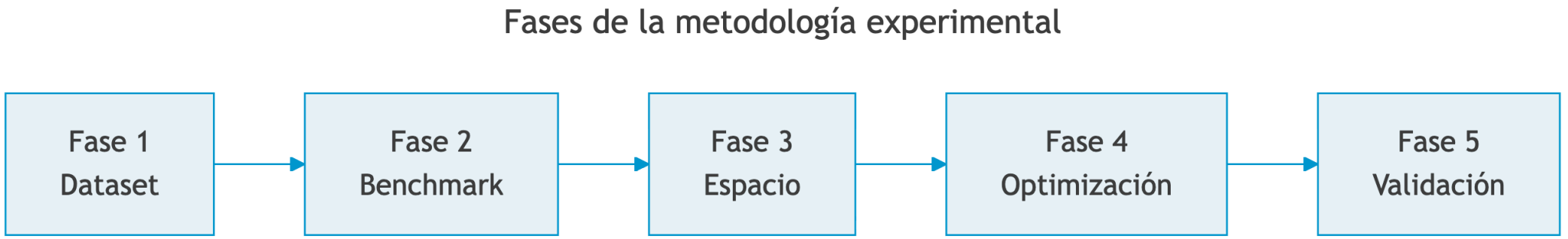
- TF y PyTorch
- Soporte español limitado
- Rápido en inferencia

Pipeline de un sistema OCR moderno



Pipeline de un sistema OCR moderno

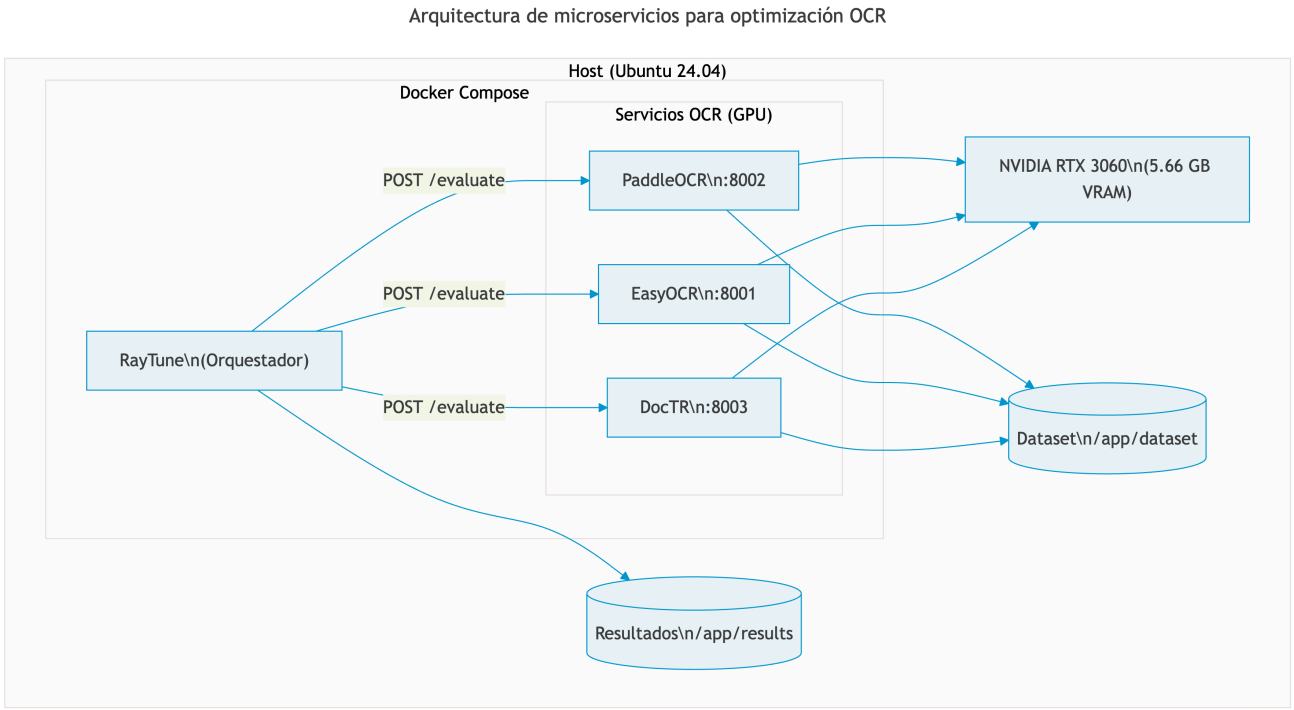
Metodología: 5 Fases



Fases de la metodología experimental

1. Preparación del dataset
   PDF → 300 DPI + GT
2. Benchmark comparativo
   3 motores, CER/WER
3. Espacio de búsqueda
   7 hiperparámetros
4. Optimización
   64 trials, TPE
5. Validación
   45 páginas completas

Arquitectura: Microservicios Docker



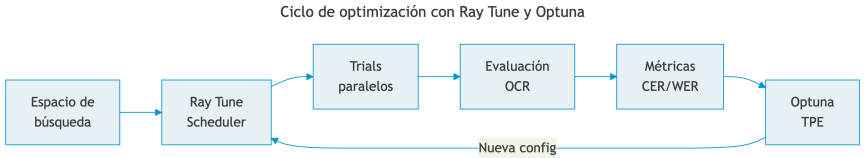
Arquitectura de microservicios para optimización OCR

- **Contenedor Ray Tune:** Orquestador de trials (Optuna TPE)
- **Contenedor OCR:** PaddleOCR con acceso GPU
- **Comunicación:** REST API (HTTP POST /evaluate)
- **Respuesta:** JSON {CER, WER, TIME}
- **Docker Compose:** Despliegue reproducible

Hardware:
RTX 3060 Laptop (5.66 GB VRAM)
AMD Ryzen 7 5800H
16 GB DDR4 | Ubuntu 24.04

Espacio de Búsqueda: 7 Hiperparámetros

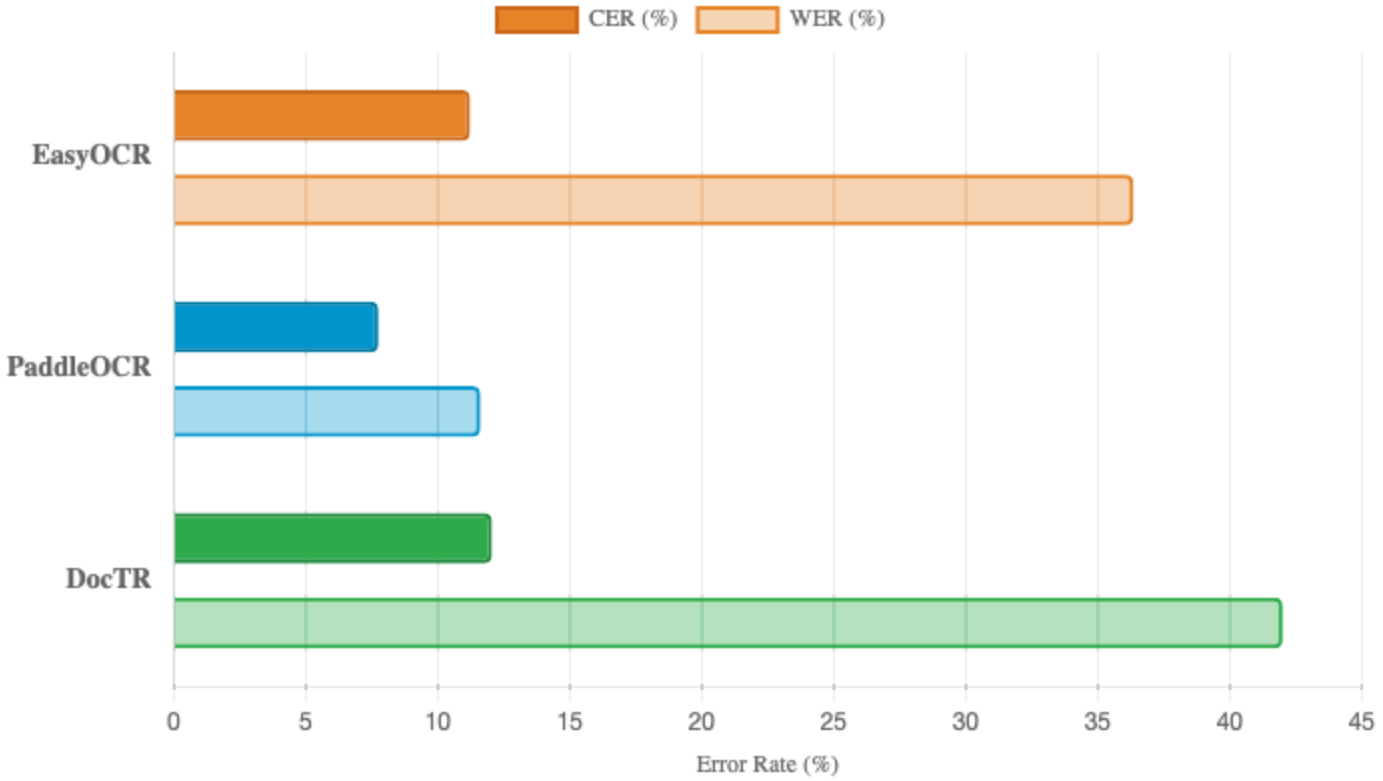
Parámetro	Tipo	Rango
`textline_orientation`	Booleano	True / False
`use_doc_orientation_classify`	Booleano	True / False
`use_doc_unwarping`	Booleano	True / False
`text_det_thresh`	Continuo	[0.01, 0.50]
`text_det_box_thresh`	Continuo	[0.01, 0.90]
`text_rec_score_thresh`	Continuo	[0.01, 0.99]
`text_det_unclip_ratio`	Fijo	0.0



Ciclo de optimización con Ray Tune y Optuna

Algoritmo: TPE (Tree-structured Parzen Estimator)
Trials: 64 | **Concurrencia:** 2 workers
Métrica: Minimizar CER

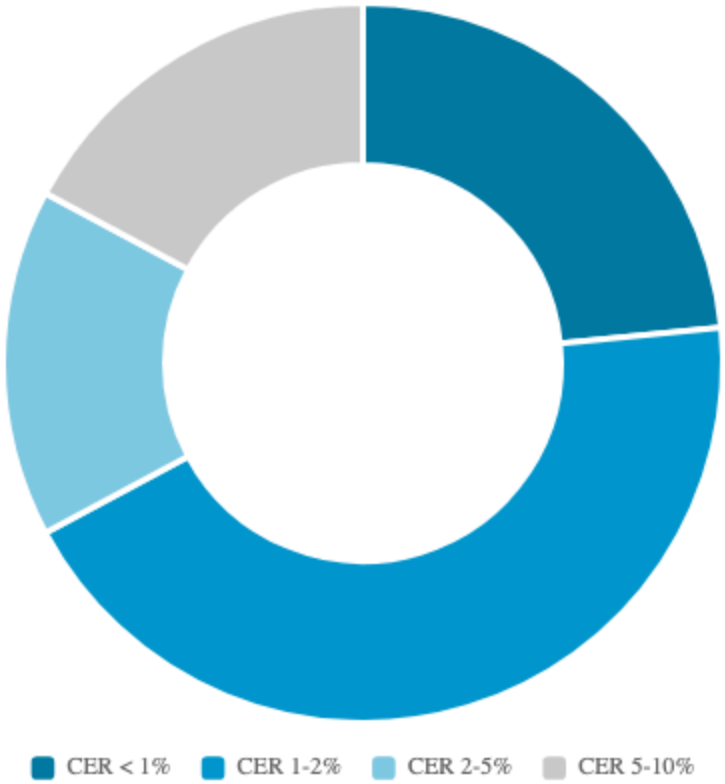
Resultados: Benchmark Comparativo



Motor	Base	HPO	Mejora
EasyOCR	11.23%	5.84%	-48%
PaddleOCR	**7.76%**	**0.79%**	**-90%**
DocTR	12.06%	7.43%	-38%

> Solo **PaddleOCR** alcanza CER<2% (43/64 trials). Mejora del **89.8%**.

Resultados: 64 Trials de Optimización



0.79%
Mejor CER (Trial #1)

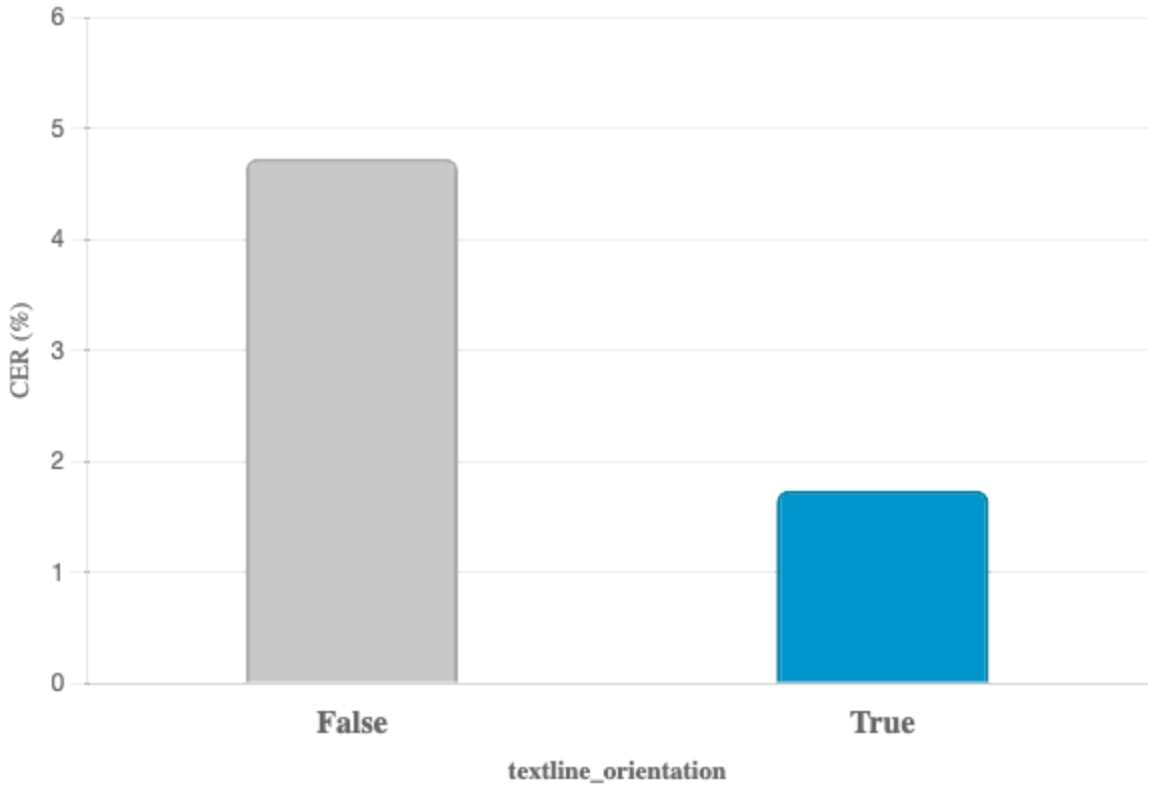
0.87%
Mediana CER

7.30%
Peor CER

67.2%
Trials con CER < 2%

0 fallos en 64 trials | Tiempo total: **~5 min** (GPU)

Hallazgo Clave: `textline_orientation`



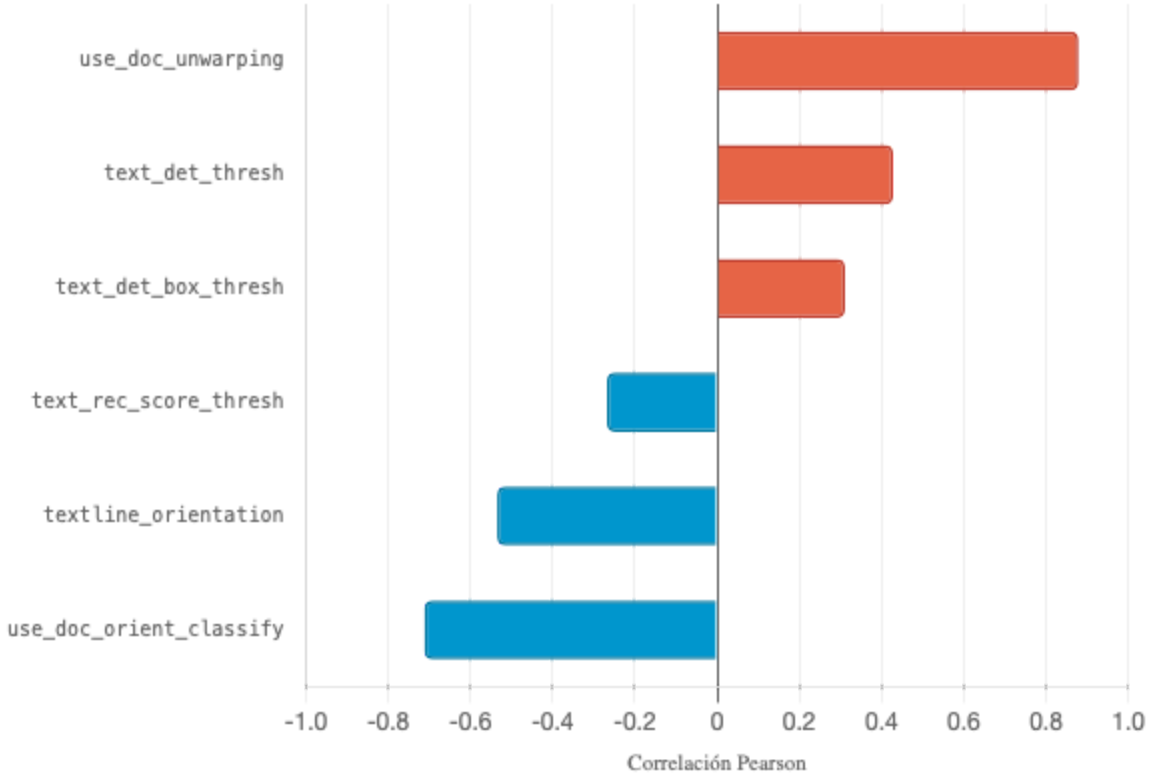
-63.2%

Reducción en CER

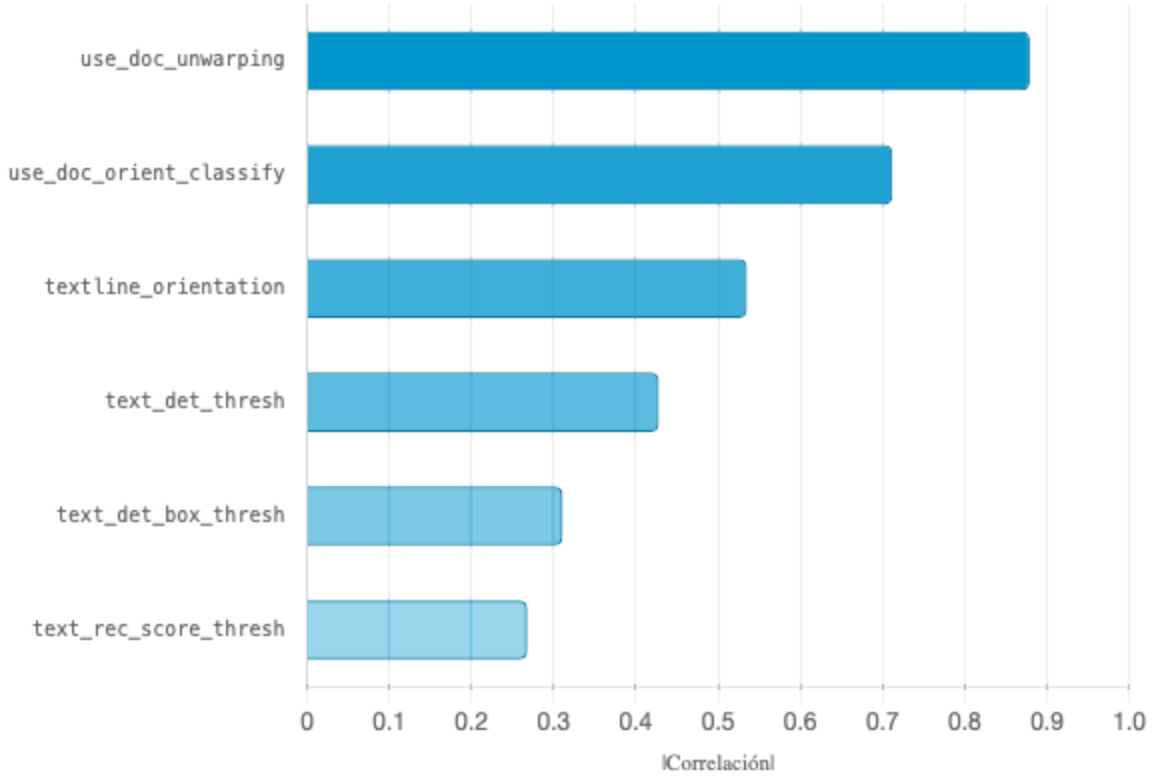
- Un **único parámetro booleano** tiene mayor impacto que todos los umbrales numéricos combinados
- **Decisiones arquitecturales** > ajustes numéricos finos
- Crítico para documentos con **layouts complejos** (índices, listas, encabezados)
- 52 de 64 trials (81%) lo activaron automáticamente (Optuna aprendió rápido)

Análisis de Hiperparámetros

Correlación Pearson con CER

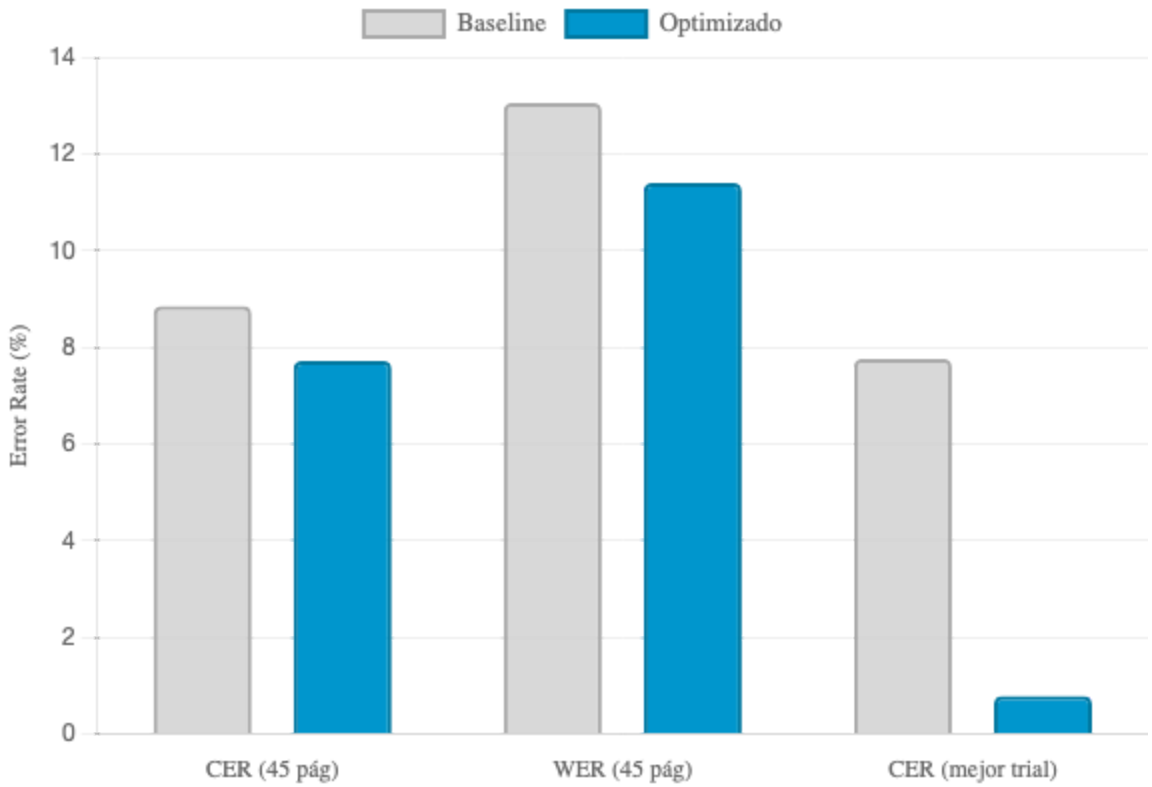


Importancia de Hiperparámetros



Insight: `use_doc_unwarping` (+0.88) es perjudicial en PDFs digitales (añade procesamiento innecesario). Los parámetros booleanos (arquitecturales) dominan sobre los umbrales numéricos.

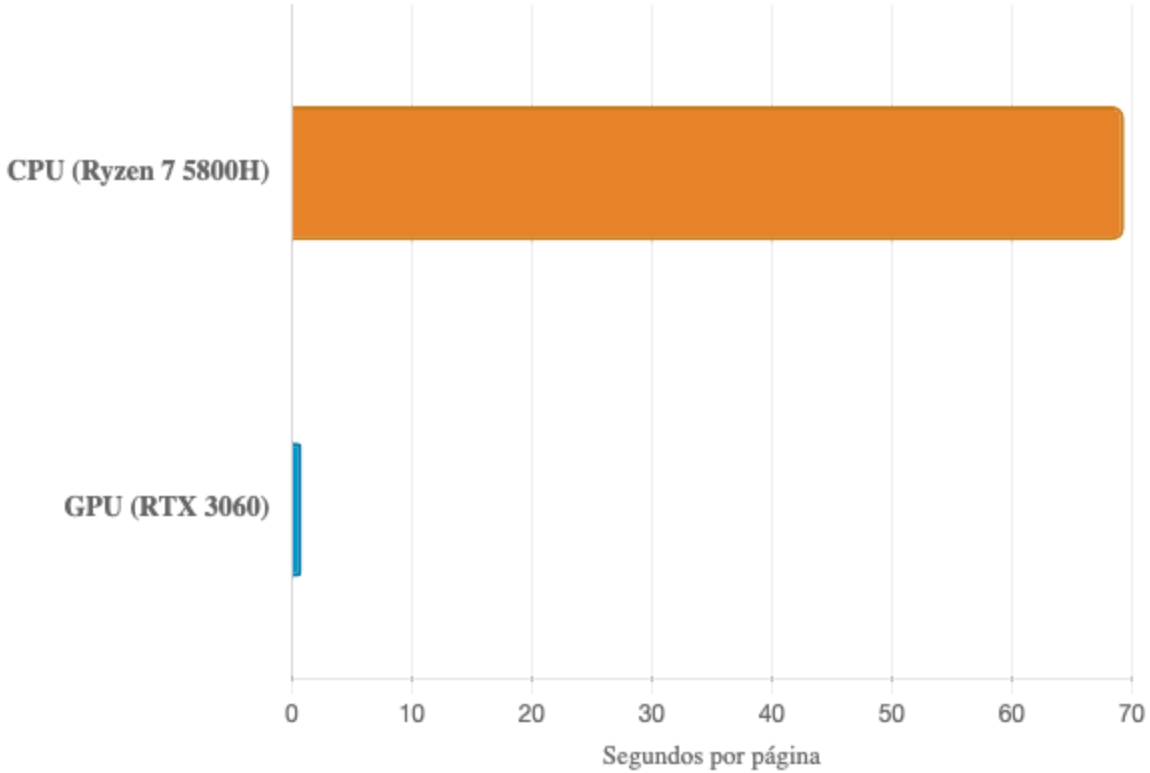
Validación: Baseline vs Optimizado



Métrica	Baseline	Optimizado	Mejora
CER (45 pág)	8.85%	7.72%	-12.8%
WER (45 pág)	13.05%	11.40%	-12.6%
CER (mejor trial, 5 pág)	**7.76%**	**0.79%**	-89.8%

Nota: La diferencia entre el mejor trial (0.79%) y la validación completa (7.72%) evidencia **sobreajuste** al subconjunto de 5 páginas usado en la optimización. Un subconjunto más amplio (15-20 páginas) mejoraría la generalización.

Aceleración GPU



82x

Factor de aceleración

0.84 s

GPU: segundos por página

69.4 s

CPU: segundos por página

64 trials × 5 páginas:
CPU: ~6.2 horas
GPU: ~5 minutos

Configuración Óptima

```
config_optimizada = {
  "textline_orientation": True,  # CRÍTICO
  "use_doc_orientation_classify": True,
  "use_doc_unwarping": False,  # Innecesario
  "text_det_thresh": 0.0462,
  "text_det_box_thresh": 0.4862,
  "text_det_unclip_ratio": 0.0,
  "text_rec_score_thresh": 0.5658,
}
```

Insights clave

- `textline_orientation = True`: Parámetro más impactante (-63.2% CER)
- `use_doc_unwarping = False`: Procesamiento innecesario para PDFs digitales
- `text_det_thresh` **bajo**: Captura más regiones de texto, reduce omisiones
- **Parámetros booleanos** dominan sobre umbrales numéricos

> Esta configuración es directamente aplicable a otros documentos académicos en español con layouts similares.

Conclusiones

Contribuciones

1. **Metodología reproducible** para optimización de hiperparámetros OCR con código abierto
2. **Análisis sistemático** de hiperparámetros PaddleOCR con correlaciones Pearson
3. **Configuración validada** para documentos académicos en español (CER 0.79%)
4. **Infraestructura dockerizada** reproducible con imágenes públicas

Limitaciones

- ! Un único tipo de documento (académico UNIR)
- ! Corpus modesto (45 páginas)
- ! Sobreajuste al subconjunto de optimización (5 páginas)
- ! `text_det_unclip_ratio` no explorado

Líneas de Trabajo Futuro

Extensiones inmediatas

- Validación cruzada en otros tipos de documentos (facturas, formularios, manuscritos)
- Subconjunto de optimización más amplio (15-20 páginas)
- Exploración de `text_det_unclip_ratio`

Líneas de investigación

- Transfer learning de hiperparámetros entre dominios
- Optimización multi-objetivo (CER + WER + velocidad)
- Comparación rigurosa HPO vs fine-tuning

Aplicaciones prácticas

- Herramienta de configuración automática por tipo de documento
- Integración en pipelines de producción
- Benchmark público de OCR en español

muchas gracias.