CLASIFICACIÓN INTELIGENTE DE DOCUMENTOS

Automatizando la Identificación de Documentos Vacíos

DAVIVIENDA

19 Agosto 2024

Indice

ESTRATEGIA

EL DESAFÍO: Clasificación Manual de Documentos Ineficiente

Procesos lentos, propensos a errores y que consumen recursos valiosos

1. La clasificación manual de documentos es un proceso tedioso y que requiere mucho tiempo.
2. Existe un alto riesgo de error humano en la clasificación manual.
3. La ineficiencia en la clasificación afecta la productividad y la calidad de la información.

LA SOLUCIÓN: Visión Artificial para una Clasificación

Implementación de un modelo de Visual Transformer (ViT) para una clasificación precisa.

Se implementó un modelo ViT, una tecnología de vanguardia en Visión Artificial, para clasificar documentos como "vacíos" o "llenos" con alta precisión.

*Se exploraron inicialmente CNNs, pero se optó por ViT debido a su capacidad para generalizar con menos datos en la **nueva** metodología **innovadora** implementada..*
*Se utilizó el modelo pre-entrenado **google/vit-base-patch16-224-in21k** de Hugging Face.*
Se implementó una técnica de reducción de dimensionalidad con UMAP y clustering con HDBSCAN para el análisis de embeddings.

EL IMPACTO: Eficiencia, Precisión y Optimización de Recursos

Beneficios tangibles para el negocio y la gestión de la información.

- Automatización de la clasificación de documentos, liberando tiempo del ara tareas de mayor valor.
- Reducción significativa del error humano en la clasificación.
- Mejora en la calidad de la información y la eficiencia de los procesos.

PROCESO

1. Preparación de Datos

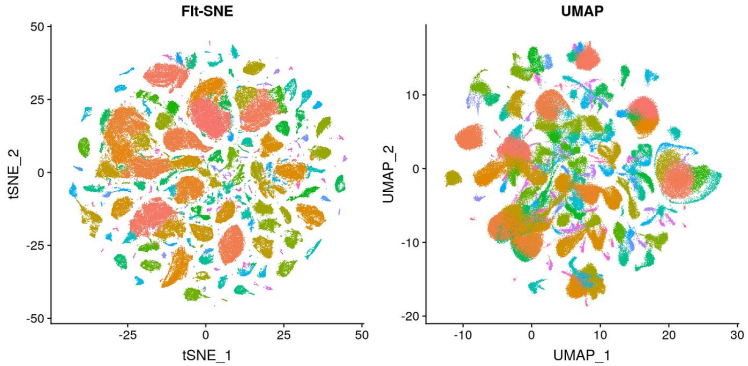
Conjunto de Datos Robusto y Balanceado

- Se utilizaron imágenes de documentos escaneados del banco y se complementaron con datos del dataset de documentos de Kaggle.
- Se balanceó el conjunto de datos para asegurar una representación equitativa de documentos "vacíos" y "llenos".
- Se incluyeron diferentes tipos de documentos para mejorar la capacidad de generalización del modelo.

2. MODELO Y EMBEDDINGS:

imágenes en representaciones vectoriales significativas.

Se cargó el modelo pre-entrenado Google / vit-base-patch16-224-in21k de Hugging Face y se generan embeddings (vectores de 768 dimensiones) para cada imagen.



Limpieza:

- *Se añadieron 299 imágenes del dataset de Kaggle (enlace).*
- *Se logró un balanceo del dataset con 276 documentos "llenos" y 270 documentos "vacíos".*

Modelo :

- *Se utilizó la librería transformers de Hugging Face para la carga del modelo y la generación de embeddings.*
- *Los embeddings representan el "significado" de la imagen y son utilizados para la clasificación.*

PROCESO

3. Reducción de Dimensionalidad y Clustering

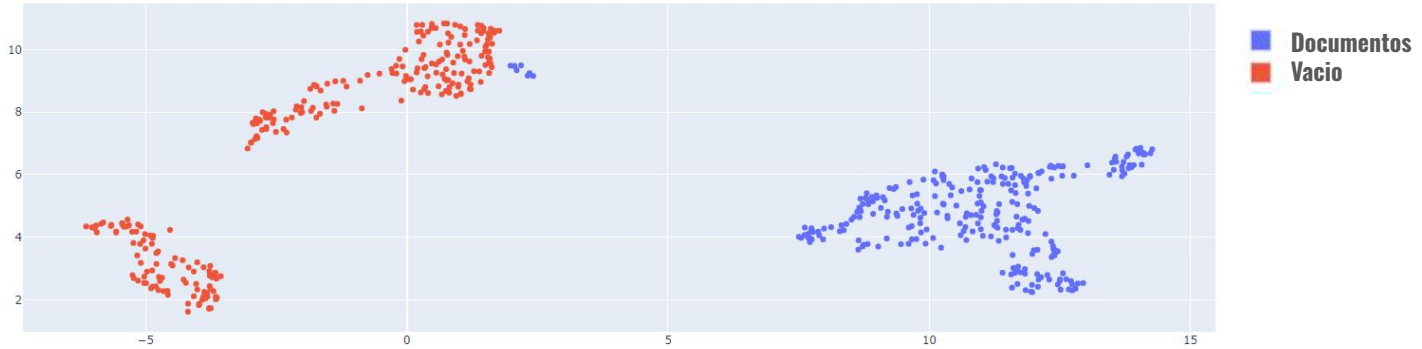
Visualizando y agrupando documentos en base a sus características.

- Se utilizó UMAP para reducir la dimensionalidad de los embeddings a 2 dimensiones.
- Se aplicó HDBSCAN para la identificación de clusters, logrando una separación clara entre documentos "vacíos" y "llenos".

4. SISTEMA DE CLASIFICACIÓN

Un enfoque eficiente y preciso para la clasificación de documentos.

Se crearon vectores representativos (promedio de embeddings) para cada cluster (vacío, lleno, vacío con información). La clasificación se realiza comparando el embedding de una nueva imagen con estos vectores representativos.



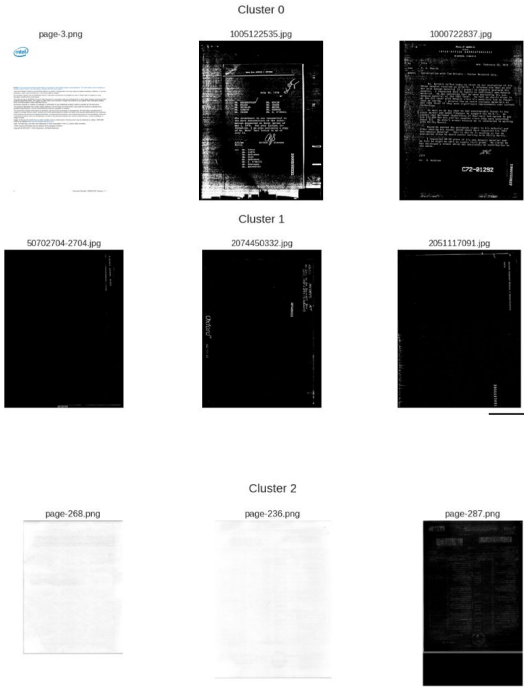
Sistema de clasificación:

- *Se utilizaron los embeddings del conjunto de entrenamiento para crear los vectores representativos.*
- *La clasificación se basa en la distancia coseno entre el embedding de la nueva imagen y los vectores representativos.*

Modelo :

- *Se utilizó la librería transformers de Hugging Face para la carga del modelo y la generación de embeddings.*
- *Los embeddings representan el "significado" de la imagen y son utilizados para la clasificación.*

RESULTADOS



***No solo** fue capaz de segmentar los documentos vacios y llenos, también los documentos vacíos **con algún tipo de información** leve*

Cluster 0

Documentos llenos

Cluster 1

Documentos vacíos con marcas

Cluster 2

Documentos vacíos o con manchas de SCAN

Comparación modelos:

- *Se utilizó una matriz de confusión para evaluar el rendimiento del modelo.*
- *Se obtuvieron resultados satisfactorios en las métricas de accuracy, precision y recall.*

Comparación modelos:

- *Se observó un buen rendimiento del modelo en la identificación de documentos "vacíos" con información adicional (ruido, marcas, etc.).*
- *Se identificaron algunos casos de clasificación errónea, principalmente debido a la calidad de las imágenes o a la presencia de elementos inusuales en los Documentos.*

RESULTADOS

EL DESAFÍO: Alta Precisión en la Clasificación de Documentos

CONJUNTO	ENTRENAMIENTO	PRUEBA
ACCURACY	98%	98%
PRECISION	96%	99%
RECALL	99%	98%

El desarrollo del modelo se realizó utilizando Google Colab y Drive,sin incurrir en costos de infraestructura, pero se debe evaluar la viabilidad de escalar el modelo a una infraestructura más robusta para producción.

El modelo ViT demostró una alta precisión en la clasificación de documentos, superando las expectativas iniciales y mostrando un gran potencial para la automatización del proceso.

Comparación modelos:
- *Se utilizó una matriz de confusión para evaluar el rendimiento del modelo.*
- *Se obtuvieron resultados satisfactorios en las métricas de accuracy, precision y recall.*

Comparación modelos:
- *Se observó un buen rendimiento del modelo en la identificación de documentos "vacíos" con información adicional (ruido, marcas, etc.).*
- *Se identificaron algunos casos de clasificación errónea, principalmente debido a la calidad de las imágenes o a la presencia de elementos inusuales en los Documentos.*

FUNCIONALIDAD

VALOR PARA EL NEGOCIO:

- **Aumento de la Eficiencia:**
  Automatización del proceso de clasificación, liberando tiempo del personal.
- **Icono de Base de Datos:**
  Mejora de la Calidad de la Información: Eliminación de errores humanos en la clasificación y optimización de la gestión de documentos.
- **Reducción de Costos:**
  Optimización de recursos y minimización de errores.

Adopción y Usabilidad:

El modelo se puede integrar fácilmente en los flujos de trabajo actuales, adaptándose a las necesidades específicas de cada área.

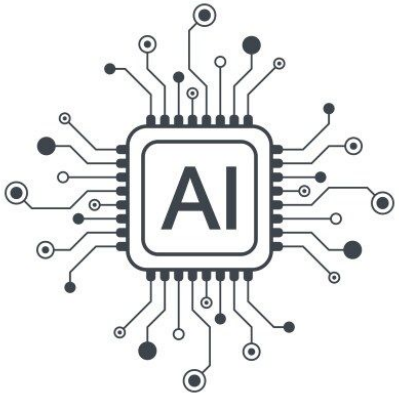
Escalabilidad:

El modelo es escalable y puede manejar grandes volúmenes de documentos, adaptándose al crecimiento del negocio.

CONCLUSIONES

Hallazgos Clave:

- Se confirmó la eficiencia de los modelos ViT para la clasificación de imágenes, incluso con conjuntos de datos relativamente pequeños.
- Se identificó la importancia de la calidad de las imágenes para el rendimiento del modelo.
- Se demostró el potencial de la automatización para la optimización de procesos y la reducción de costos.



Recomendaciones y Próximos Pasos:

- Implementar el modelo en un entorno de producción para la clasificación automatizada de documentos.
- Continuar con la optimización del modelo para mejorar su precisión y eficiencia.
- Explorar la posibilidad de aplicar el modelo a otros tipos de documentos o procesos.

Comparación modelos:

- *Se requiere una inversión en infraestructura para la implementación del modelo en producción.*
- *Se estima un retorno de la inversión a corto plazo debido a la optimización de recursos y la mejora en la eficiencia operativa.*

IMPLEMENTACION

Mejoras Potenciales:

- Implementar un sistema de control de calidad para las imágenes escaneadas.
- Entrenar el modelo con un conjunto de datos aún mayor y más diverso.
- Explorar la posibilidad de utilizar técnicas de aprendizaje por transferencia (transfer learning) para mejorar el rendimiento del modelo.

Viabilidad Técnica y Económica:

○Viabilidad Técnica: El modelo es técnicamente viable y se puede implementar con la infraestructura adecuada.
○ Viabilidad Económica: El proyecto ofrece un alto potencial de retorno de la inversión, con una reducción de costos y una mejora en la eficiencia operativa.

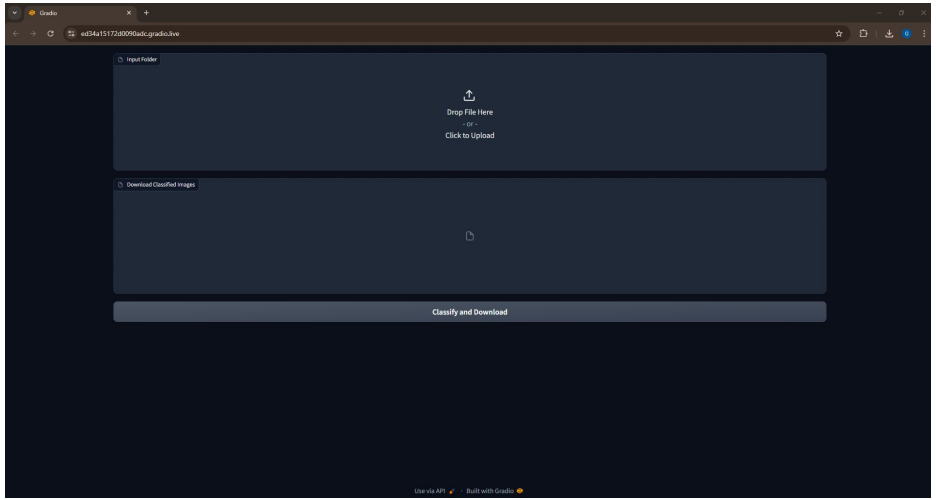
Consideraciones Futuras:

○ Monitorear el rendimiento del modelo en producción y reentrenarlo periódicamente con nuevos datos.
○ Adaptar el modelo a las nuevas necesidades del negocio, como la clasificación de diferentes tipos de documentos.
○ Investigar nuevas tecnologías y modelos de Visión Artificial para mejorar la precisión y la eficiencia del sistema.

Mejoras:
○ Se pueden utilizar técnicas de aumento de datos (data augmentation) para mejorar la generalización del modelo.
○ Se pueden explorar otras arquitecturas de ViT o modelos de Visión Artificial para comparar su rendimiento.

API

El despliegue del modelo se realizó con Gradio [VIDEO]



Algunos flujos del funcionamiento de la API:

Flujo de datos