Titulo

En

Donut Forget the Details: OCR-Free end-to-end Transformer-Powered Visual Document Understanding Model.

Es

A Donut no le falta detalle: Modelo end-to-end para comprensión visual de documentos basado en Transformers sin OCR.

Val

A Donut no li falta detall: Model end-to-end per a la comprensió visual de documents basat en Transformers sense OCR.

Resumen

En

Document processing has traditionally relied on the use of Optical Character Recognition (ORC) to extract text from a document or image. This text is then fed into a Large Language Model (LLM), such as the well-known Chat-GPT, which understands the information and produces a structured output.

However, OCR-based approaches often struggle with complex layouts, handwritten text, and noisy document conditions, leading to loss of contextual information and inaccuracies. In addition, OCR pipelines are very expensive and cannot be run on with standard computers. What we propose is an ORC-free end-to-end solution for Visual Document Understanding tasks: Donut.

Donut is a Transformer-based architecture model which processes documents as images and combines a vision encoder with a sequence-to-sequence text decoder to process the documents and provide an structured output in the desired format. By eliminating OCR as an intermediate step, Donut can better preserve the spatial and semantic relationships within documents, and does not pass on any of the potential errors of ORC to the LLM, resulting in improved performance for key information extraction, document classification, and question-answering tasks.

The research will focus on evaluating Donut's performance against state of the art options, to consider which solution is better. To achieve this, we will focus on metrics and how much data each model needs to be fine-tuned for a specific task.

Es

El tratamiento de documentos se ha basado tradicionalmente en el uso del reconocimiento óptico de caracteres (ORC) para extraer texto de un documento o imagen. A continuación, este texto se introduce en un Gran modelo de lenguaje (LLM), como lo es Chat-GPT, que comprende la información y produce una salida estructurada.

Sin embargo, los enfoques basados en OCR suelen tener dificultades con diseños complejos, texto manuscrito y documentos ruidosos, lo que provoca pérdidas de información contextual e

imprecisiones. Además, los OCR son muy costoso computacionalmente y no pueden ejecutarse en ordenadores normales: hay que usar servicios es la nube. Lo que proponemos es una solución 'end-to-end' sin OCR para tareas de comprensión visual de documentos: Donut.

Donut es un modelo con arquitectura basada en 'Transformers' que procesa los documentos como imágenes y combina un 'vision enconcer' con un 'text-decoder' para procesar los documentos y proporcionar una salida estructurada en el formato deseado. Al eliminar el OCR como paso intermedio, Donut puede preservar mejor las relaciones espaciales y semánticas dentro de los documentos, y no transmite ninguno de los errores potenciales del ORC al LLM. Esto se traduce en un rendimiento mejorado para tareas clave de extracción de información, clasificación de documentos y respuesta a preguntas.

La investigación se centrará en evaluar el rendimiento de Donut frente a las opciones actuales o 'state of the art', para considerar qué solución es la más apropiada. Para ello, nos centraremos en las métricas y en la cantidad de datos que necesita cada modelo para afinarse en una tarea específica.

Val

El tractament de documents s'ha basat tradicionalment en l'ús del reconeixement òptic de caràcters (ORC) per a extraure text d'un document o imatge. A continuació, aquest text s'introdueix en un Gran model de llenguatge (LLM) , un exemple seria Chat-GPT, que comprén la informació i produeix una eixida estructurada.

No obstant això, els enfocaments basats en OCR solen tindre dificultats amb dissenys complexos, text manuscrit i documents sorollosos, la qual cosa provoca pèrdues d'informació contextual i imprecisions. A més, els OCR són molt costós computacionalment i no poden executar-se en ordinadors normals: cal utilitzar servicis en el núvol. El que proposem és una solució 'end-to-end' sense OCR per a tasques de comprensió visual de documents: Donut.

Donut és un model amb arquitectura basada en 'Transformers' que processa els documents com a imatges i combina un '*vision enconcer' amb un 'text-decoder' per a processar els documents i proporcionar una eixida estructurada en el format desitjat. En eliminar el OCR com a pas intermedi, Donut pot preservar millor les relacions espacials i semàntiques dins dels documents, i no transmet cap dels errors potencials del ORC al LLM. Això es tradueix en un rendiment millorat per a tasques clau d'extracció d'informació, classificació de documents i resposta a preguntes.

La investigació se centrarà en avaluar el rendiment de Donut enfront de les opcions actuals o 'state of the art', per a considerar quina solució és la més apropiada. Per a això, ens centrarem en les mètriques i en la quantitat de dades que necessita cada model per a afinar-se en una tasca específica.

Keywords

- Visual Document Understanding
- Document Processing
- OCR-Free
- Donut Model
- End-to-End