OCR-free Document Understanding Transformer (Donut)

G. Kim, T. Hong, M. Yim NAVER CLOVA AI Lab

Clark University | Fall 2025 | Professor: Dr. Onur Barut

- Charan D

Reason to select

- It introduces a new methodology for document AI — removing OCR and training a fully generative Transformer model for structured document understanding.

Background

Domain: Visual Document Understanding (VDU) – extracting structured data from document images (receipts, forms, invoices).

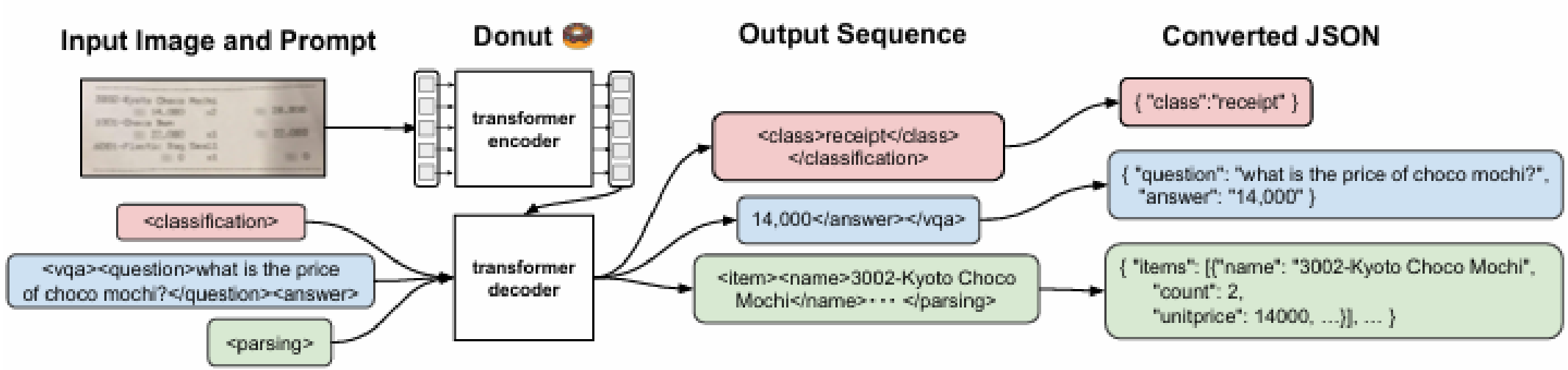
Traditional approach: OCR → Text Parsing → Information Extraction.

Problem: OCR pipelines = slow, language-specific, error-propagating, expensive.

Need: A unified end-to-end model that learns directly from pixels to text.

Summary

- Proposes **Donut**, the first OCR-free Transformer for VDU.
- End-to-end training on raw document images → structured JSON outputs.
- Introduces **SynthDoG**, a synthetic document generator for multi-lingual pre-training.
- Achieves **state-of-the-art** accuracy and 2× speed vs OCR-dependent models.
- Open-sourced model & data for reproducibility.



Methodology Overview

Vision-Language Encoder–Decoder Transformer

- **Encoder:** Swin Transformer extracts visual patch features.
- **Decoder:** BART-style Transformer generates token sequence (JSON/text).
- **Training objective:** Cross-entropy next-token prediction.
- **Task prompting:** JSON fields or question-answer tokens (e.g., `<s_cord-v2>`).

Model Details

Input: Document image (2560 × 1920 pixels).

Encoder output: Patch embeddings → latent vectors.

Decoder input: Prompt tokens + image embeddings.

Output format: Structured JSON using special tokens [START *] [END *].

Parameters: ≈ 220 M (total 143 M trainable + OCR-free architecture).

Languages: English, Korean, Japanese, Chinese

Setup

Benchmarks: RVL-CDIP (document classification), CORD (receipts), Ticket (Chinese train tickets), Business Cards (Japanese), Receipts (Korean), DocVQA.

Metrics: Accuracy, F1 score, Tree Edit Distance (TED), ANLS for QA.

Baselines: BERT, LayoutLM, LayoutLMv2, BROS, WYVERN, SPADE.

Hardware: Trained on 64 × A100 GPUs (200 K steps).

Results

Task	Model	Accuracy / F1	Speed (s/img)
Classification (RVL-CDIP)	Donut 95.3%	0.75	Faster & better than LayoutLMv2 (95.25%)
CORD IE	Donut 84.1 F1 / 90.9 Acc	1.2	+12% vs BERT baseline
Ticket IE	Donut 94.1 F1	0.6	High accuracy Chinese domain
DocVQA	Donut ANLS 67.5	0.78	Comparable without OCR

Output

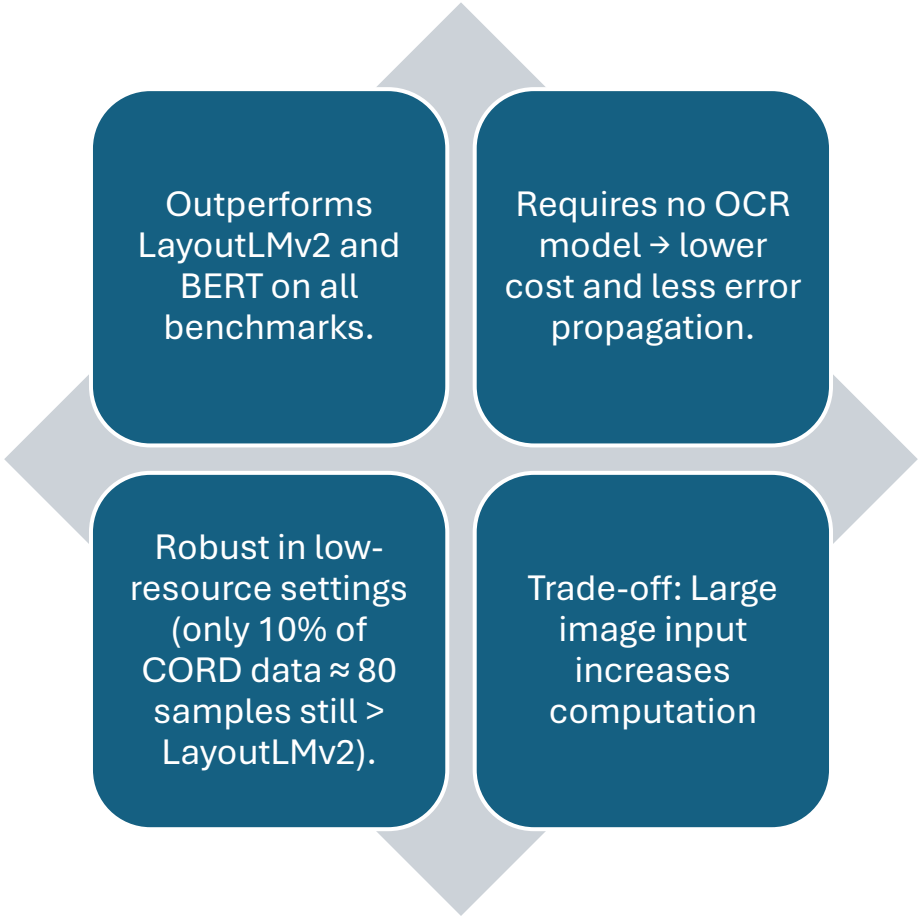
Donut generates structured JSON outputs like:

{"store":"Starbucks","items":[{"name":"Latte","price":"$4.50"}],"total":"$7.25"}

Visual attention maps show decoder focuses on text regions without OCR guidance.

Handles handwritten and multi-lingual inputs robustly

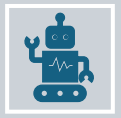
Comparison


Outperforms LayoutLMv2 and BERT on all benchmarks.
Requires no OCR model → lower cost and less error propagation.
Robust in low-resource settings (only 10% of CORD data ≈ 80 samples still > LayoutLMv2).
Trade-off: Large image input increases computation

Strengths & Limitations

Strengths	End-to-end OCR-free pipeline (simpler & faster).	Multi-lingual pre-training via SynthDoG.	Reproducible (open code + data).
Limitations	Performance drops on tiny texts / high-resolution docs.	High GPU cost for large input sizes.	JSON format rigidity – fails if token structure breaks.

Scope



Applications: Automated receipt parsing, invoice processing, KYC ID validation, and accessibility tools.

Impact: Reduces OCR cost and language bias in document AI.

Follow-ups: Donut v2 (Synthetic data expansion), DocFormer, LayoutLMv3, VisionLLMs (e.g., Pix2Struct).

Trend: Shift toward OCR-free, multimodal generative AI systems.

Takeaways

01

DONUT PROVES OCR-FREE TRANSFORMERS CAN OUTPERFORM OCR-BASED SYSTEMS.

02

SYNTHDOG ENABLES MULTI-LINGUAL AND COST-EFFECTIVE PRE-TRAINING.

03

ARCHITECTURE SIMPLICITY → INDUSTRIAL DEPLOYABILITY.

04

POINTS TO FUTURE MULTI-MODAL LLMS FOR DOCUMENT UNDERSTANDING.