# Visual Document Understanding

Martin Geletka, 456576

# Outline

Intro & Classical approaches

LayoutLM

LiT

TrOCR

Donut, SWIN Transformer

# Intro & Classical approaches

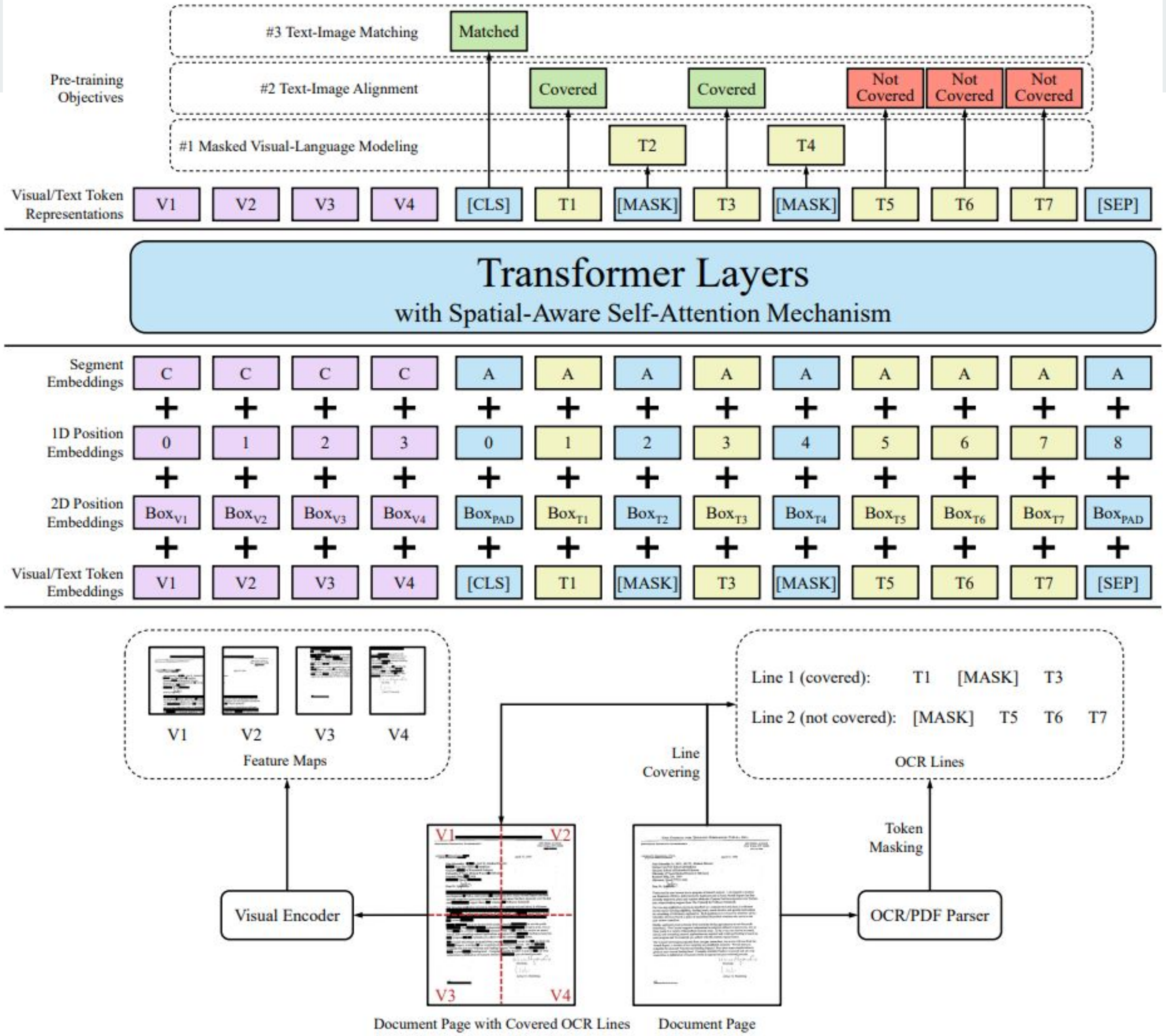
# Problems

- → OCR
- → Classification
- → NER
- → Example of use: Intelligent Back Office

# Classical approaches

- Classification
  - Connect outputs from independent NN for vision and text
  - Shallow model on top, simple confidence
- NER
  - Preprocess the document with OCR
  - Use NER model only on text data {output from OCR)

# LayoutLM


Pre-training Objectives
#3 Text-Image Matching
Matched
#2 Text-Image Alignment
Covered
Covered
Not Covered
Not Covered
Not Covered
#1 Masked Visual-Language Modeling
T2
T4
Visual/Text Token Representations
V1 V2 V3 V4 [CLS] T1 [MASK] T3 [MASK] T5 T6 T7 [SEP]
Transformer Layers
with Spatial-Aware Self-Attention Mechanism
Segment Embeddings
C C C C A A A A A A A A A
1D Position Embeddings
0 1 2 3 0 1 2 3 4 5 6 7 8
2D Position Embeddings
Box$_{V1}$ Box$_{V2}$ Box$_{V3}$ Box$_{V4}$ Box$_{PAD}$ Box$_{T1}$ Box$_{T2}$ Box$_{T3}$ Box$_{T4}$ Box$_{T5}$ Box$_{T6}$ Box$_{T7}$ Box$_{PAD}$
Visual/Text Token Embeddings
V1 V2 V3 V4 [CLS] T1 [MASK] T3 [MASK] T5 T6 T7 [SEP]
V1 V2 V3 V4
Feature Maps
Line 1 (covered): T1 [MASK] T3
Line 2 (not covered): [MASK] T5 T6 T7
OCR Lines
Line Covering
Token Masking
Visual Encoder
OCR/PDF Parser
V1 V2 V3 V4
Document Page with Covered OCR Lines
Document Page

# LayoutLM - Text Embeddings

- Preprocessing
  - WordPiece tokenizer,
  - [CLS] at the beginning of the sequence
  - [SEP] at the end of each text segment
- Final text embedding
  - Token embedding
  - Token index
  - Segment index

$$\mathbf{t}_i = \mathrm{TokEmb}(w_i) + \mathrm{PosEmb1D}(i) + \mathrm{SegEmb}(s_i)$$

# LayoutLM - Visual Embeddings

- Use pretrained ResNeXt-FPN backbone
- Pipeline
  - resized to 224 × 224
  - Fed to backbone
  - Output in size WxH
  - linear projection Ito obtain same dimensionality as text embeddings

$$\begin{aligned}\mathbf{v}_i = &\operatorname{Proj}\big(\operatorname{VisTokEmb}(I)_i\big)\\ &+\operatorname{PosEmb1D}(i)+\operatorname{SegEmb}(\texttt{[C]})\end{aligned}$$

# LayoutLM - Layout Embedding

- ➔ represent spatial layout information
- ➔ Preprocessing:
  - ◆ normalize and discretize all coordinates to integer

$$\mathbf{l}_i = \mathrm{Concat}\big(\mathrm{PosEmb2D_x}(x_{\mathrm{min}}, x_{\mathrm{max}}, width), \mathrm{PosEmb2D_y}(y_{\mathrm{min}}, y_{\mathrm{max}}, height)\big)$$

# LayoutLM - pretraining tasks

- Masked Visual-Language Modeling
  - mask some text tokens and corresponding image regions
  - The layout embedding remain
- Text Image alignment
  - Covered visual parts and classified text to Covered  vs UnCovered
- Text-Image Matching
  - Classify if text and image are from same document

# LayoutLM - Data

- Training
  - IIT-CDIP Test Collection
  - 7M documents, 40M pages, 1.5 TB
- Downstream tasks
  - Entity extraction tasks - FUNSD, CORD, SROIE, KleisterNDA
  - Document classification: RVL-CDIP,
  - QA: DocVQA

# LayoutLM - Results

| Model | Accuracy |
|---|---|
| $\text{BERT}_{\text{BASE}}$ | 89.81% |
| $\text{UniLMv2}_{\text{BASE}}$ | 90.06% |
| $\text{BERT}_{\text{LARGE}}$ | 89.92% |
| $\text{UniLMv2}_{\text{LARGE}}$ | 90.20% |
| $\text{LayoutLM}_{\text{BASE}}$ (w/ image) | 94.42% |
| $\text{LayoutLM}_{\text{LARGE}}$ (w/ image) | 94.43% |
| $\text{LayoutLMv2}_{\text{BASE}}$ | 95.25% |
| $\text{LayoutLMv2}_{\text{LARGE}}$ | **95.64%** |
| VGG-16 (Afzal et al., 2017) | 90.97% |
| Single model (Das et al., 2018) | 91.11% |
| Ensemble (Das et al., 2018) | 92.21% |
| InceptionResNetV2 (Szegedy et al., 2017) | 92.63% |
| LadderNet (Sarkhel and Nandi, 2019) | 92.77% |
| Single model (Dauphinee et al., 2019) | 93.03% |
| Ensemble (Dauphinee et al., 2019) | 93.07% |

# LayoutLM - Results

| Model | Accuracy |
|---|---|
| $\text{BERT}_{\text{BASE}}$ | 89.81% |
| $\text{UniLMv2}_{\text{BASE}}$ | 90.06% |
| $\text{BERT}_{\text{LARGE}}$ | 89.92% |
| $\text{UniLMv2}_{\text{LARGE}}$ | 90.20% |
| $\text{LayoutLM}_{\text{BASE}}$ (w/ image) | 94.42% |
| $\text{LayoutLM}_{\text{LARGE}}$ (w/ image) | 94.43% |
| $\text{LayoutLMv2}_{\text{BASE}}$ | 95.25% |
| $\text{LayoutLMv2}_{\text{LARGE}}$ | **95.64%** |
| VGG-16 (Afzal et al., 2017) | 90.97% |
| Single model (Das et al., 2018) | 91.11% |
| Ensemble (Das et al., 2018) | 92.21% |
| InceptionResNetV2 (Szegedy et al., 2017) | 92.63% |
| LadderNet (Sarkhel and Nandi, 2019) | 92.77% |
| Single model (Dauphinee et al., 2019) | 93.03% |
| Ensemble (Dauphinee et al., 2019) | 93.07% |

Table 3: Classification accuracy on the RVL-CDIP dataset

| Model | Fine-tuning set | ANLS |
|---|---|---|
| $\text{BERT}_{\text{BASE}}$ | train | 0.6354 |
| $\text{UniLMv2}_{\text{BASE}}$ | train | 0.7134 |
| $\text{BERT}_{\text{LARGE}}$ | train | 0.6768 |
| $\text{UniLMv2}_{\text{LARGE}}$ | train | 0.7709 |
| $\text{LayoutLM}_{\text{BASE}}$ | train | 0.6979 |
| $\text{LayoutLM}_{\text{LARGE}}$ | train | 0.7259 |
| $\text{LayoutLMv2}_{\text{BASE}}$ | train | 0.7808 |
| $\text{LayoutLMv2}_{\text{LARGE}}$ | train | 0.8348 |
| $\text{LayoutLMv2}_{\text{LARGE}}$ | train + dev | 0.8529 |
| $\text{LayoutLMv2}_{\text{LARGE}}$ + QG | train + dev | **0.8672** |
| Top-1 (30 models ensemble) on DocVQA Leaderboard (until 2020-12-24) | - | 0.8506 |

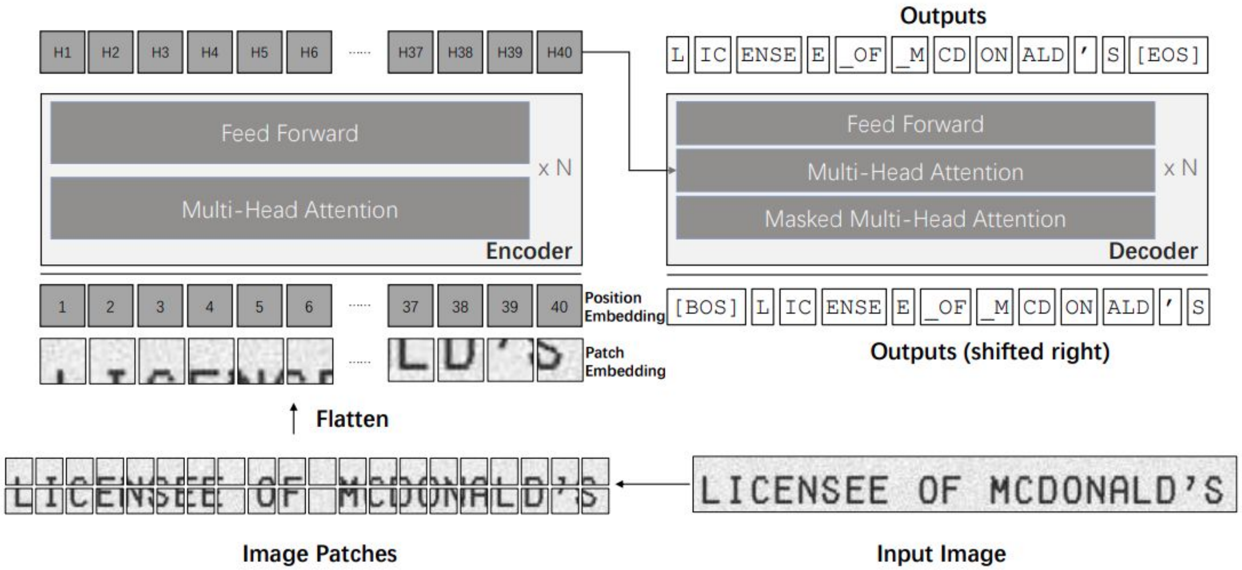
Table 4: ANLS score on the DocVQA dataset, "QG" denotes the data augmentation with the question generation dataset.

# TrOCR

# TrOCR

- Transformer based Optical Character Recognition
- Encoder - Decoder architecture
- Uses pretrained models
    - Encoder - Vision Transformer
    - Decoder - Text Transformer

# TrOCR



Outputs
H1 H2 H3 H4 H5 H6 ...... H37 H38 H39 H40
L IC ENSE E _OF _M CD ON ALD ' S [EOS]
Feed Forward
Multi-Head Attention
x N
Encoder
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
x N
Decoder
1 2 3 4 5 6 ...... 37 38 39 40
Position Embedding
Patch Embedding
[BOS] L IC ENSE E _OF _M CD ON ALD ' S
Outputs (shifted right)
Flatten
LICENSEE OF MCDONALD'S
Image Patches
Input Image

# TrOCR - training

- Pretrained on  text recognition
    - Two stages
        - Synthetically generated from text
        - On printed, handwritten data
    - Data augmentation
        - random rotation (-10 to 10 degrees), Gaussian blurring, image dilation, image erosion, downscaling, underlining or keeping the original.

# TrOCR - results

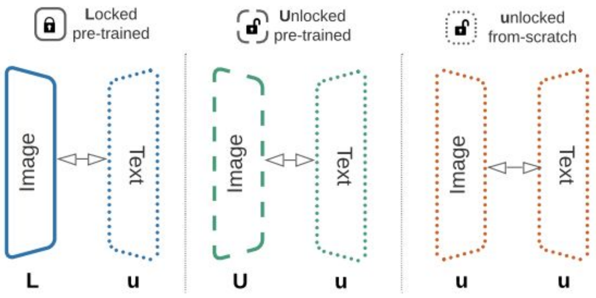
| Model | Architecture | Training Data | External LM | CER |
|---|---|---|---|---|
| $\text{TrOCR}_{\text{BASE}}$ | Transformer | Synthetic + IAM | No | 3.42 |
| $\text{TrOCR}_{\text{LARGE}}$ | Transformer | Synthetic + IAM | No | 2.89 |
| (Bluche and Messina, 2017) | GCRNN / CTC | Synthetic + IAM | Yes | 3.2 |
| (Michael et al., 2019) | LSTM/LSTM w/Attn | IAM | No | 4.87 |
| (Wang et al., 2020) | FCN / GRU | IAM | No | 6.4 |
| (Kang et al., 2020) | Transformer w/ CNN | Synthetic + IAM | No | 4.67 |
| (Diaz et al., 2021) | S-Attn / CTC | Internal + IAM | No | 3.53 |
| (Diaz et al., 2021) | S-Attn / CTC | Internal + IAM | Yes | 2.75 |
| (Diaz et al., 2021) | Transformer w/ CNN | Internal + IAM | No | 2.96 |

Table 4: Evaluation results (CER) on the IAM Handwriting dataset.

# LiT

# LiT - Locked-image Tuning

- ➔ Contrastive training
  - ◆ Goal:
    - representations of paired images and texts to be similar
    - representations of non-paired images and texts to be dissimilar
- ➔ Locked image Tuning
  - ◆ Locked image/text pretrained embeddings and move the others



# LiT - Results

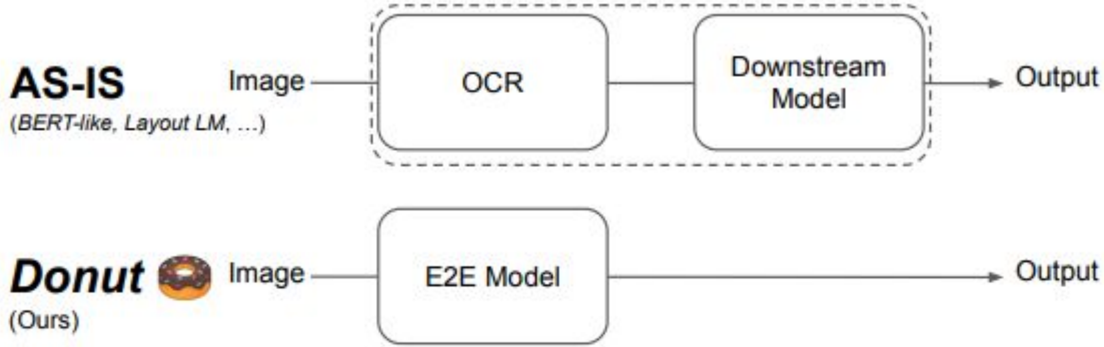
- Datasets
  - CC12M
  - YFCC100m

| Dataset | Method | INet | INet-v2 | INet-R | INet-A | ObjNet | ReaL | VTAB-N |
|---|---|---|---|---|---|---|---|---|
| Private | CLIP [45] | 76.2 | 70.1 | 88.9 | 77.2 | 72.3 | - | - |
| | ALIGN [30] | 76.4 | 70.1 | 92.2 | 75.8 | - | - | - |
| | *LiT* | **84.5** | **78.7** | **93.9** | **79.4** | **81.1** | 88.0 | 72.6 |
| Public | CLIP [45] | 31.3 | - | - | - | - | - | - |
| | OpenCLIP [28] | 34.8 | 30.0 | - | - | - | - | - |
| | *LiT* | **75.7** | **66.6** | 60.4 | 37.8 | 54.5 | 82.1 | 63.1 |
| * | ResNet50 [25] | 75.8 | 63.8 | 36.1 | 0.5 | 26.5 | 82.5 | 72.6 |

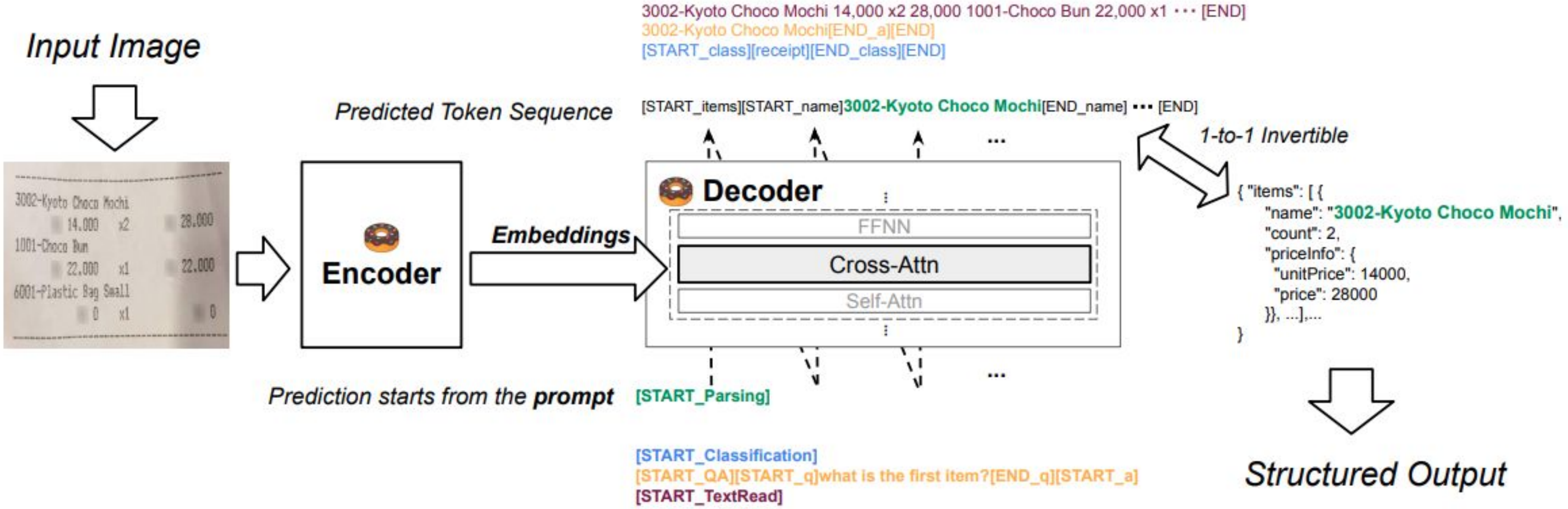
# Donut

# Donut - Idea

➔ Document Understanding Transformer without OCR

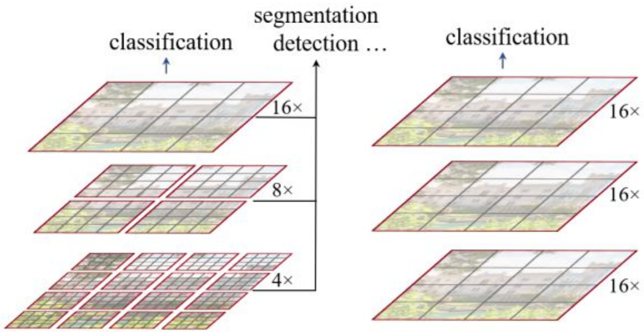


# Donut - architecture



# Donut - Encoder SWIN Transformer

- ➔ Two main ideas
  - ◆ Hierarchical - better represents small regions
  - ◆ Shifted windows

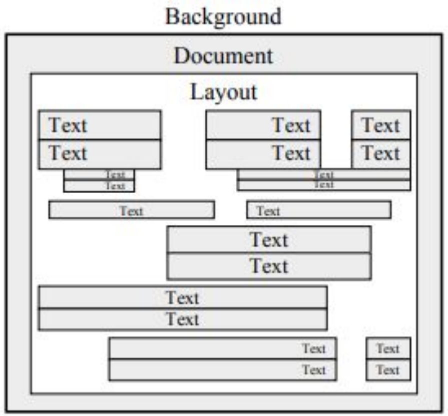



(a) Swin Transformer (ours)

(b) ViT

# SynthDoG

- ➔ Synthetic Document Generator
- ➔ Pipeline
    - ◆ Background - sample from ImageNet
    - ◆ Texture - sampled from collected photos
    - ◆ Words - sampled from Wikipedia
    - ◆ Patterns -rule based random patterns




# Donut - pretraining

➔ generated 1.2M synthetic document images

➔ model is trained to read all the texts in the images in the reading order from top left to bottom right

# Donut - Downstream tasks

- ➔ Document Classification - RVLCDIP
- ➔ Document Parsing - Indonesian Receipts, Japanese Business Cards, Korean Receipts
- ➔ Document VQA

# Donut - Results - Classification

| | use OCR | #Params | Time(ms) | Accuracy (%) |
|---|---|---|---|---|
| $\text{BERT}_{\text{BASE}}$ | ✓ | 110M + n/a[†] | 1392 | 89.81 |
| $\text{RoBERTa}_{\text{BASE}}$ | ✓ | 125M + n/a[†] | 1392 | 90.06 |
| $\text{UniLMv2}_{\text{BASE}}$ | ✓ | 125M + n/a[†] | n/a | 90.06 |
| $\text{LayoutLM}_{\text{BASE}}$ (w/ image) | ✓ | 160M + n/a[†] | n/a | 94.42 |
| $\text{LayoutLMv2}_{\text{BASE}}$ | ✓ | 200M + n/a[†] | 1489 | **95.25** |
| **Donut (Proposed)** | | 156M | **791** | <u>94.50</u> |

[†] Parameters for OCR should be considered for the non-E2E models.

# Donut - Results - Document Parsing

| | use OCR | Params | Indonesian Receipt | | Korean Receipt | | Japanese Business Card | |
|---|---|---|---|---|---|---|---|---|
| | | | Time (s) | nTED | Time (s) | nTED | Time (s) | nTED |
| BERT-based Extractor* | ✓ | 86M† + n/a‡ | 0.89 + 0.54 | 11.3 | 1.14 + 1.74 | 21.67 | 0.83 + 0.50 | 9.56 |
| SPADE (Hwang et al., 2021b) | ✓ | 93M† + n/a‡ | 3.32 + 0.54 | 10.0 | 6.56 + 1.74 | 21.65 | 3.34 + 0.50 | 9.77 |
| **Donut (Proposed)** | | 156M† | 1.07 | **8.45** | 1.99 | **5.87** | 1.39 | **3.70** |

# Donut - Results - DocVQA

| | OCR | Params[‡] | Time (ms) | ANLS |
|---|---|---|---|---|
| LoRRA | ✓ | ~223M | n/a | 11.2 |
| M4C | ✓ | ~91M | n/a | 39.1 |
| $\text{BERT}_{\text{BASE}}$ | ✓ | 110M | n/a | 57.4 |
| CLOVA OCR | ✓ | n/a | $\gtrsim$ 3226 | 32.96 |
| UGLIFT v0.1 | ✓ | n/a | $\gtrsim$ 3226 | 44.17 |
| $\text{BERT}_{\text{BASE}}$ | ✓ | 110M + n/a[†] | 1517 | 63.54 |
| $\text{LayoutLM}_{\text{BASE}}$ | ✓ | 113M + n/a | 1519 | 69.79 |
| $\text{LayoutLMv2}_{\text{BASE}}$ | ✓ | 200M + n/a | 1610 | 78.08 |
| **Donut** | | ~207M | 809 | 47.14 |
| + 10K imgs of trainset | | | | 53.14 |

# Questions

# References


- LayoutLM v2 -> https://arxiv.org/pdf/2012.14740.pdf
- TrOCR -> https://arxiv.org/pdf/2109.10282.pdf
- LiT -> https://arxiv.org/abs/2111.07991
- Donut -> https://arxiv.org/pdf/2111.15664.pdf
- SWIN Transformer -> https://arxiv.org/pdf/2103.14030.pdf