Document AI

• Analyze forms and documents

• Create intelligent search indexes

• Automate business workflows

Uncover latent insights from all your content
Document AI in Real World

**Form**

Scanned documents (.jpg, .png, ...)

**Receipt**

**Report**

Digital-born documents (.pdf, .docx, ...)

**Invoice**

*Layout invariance* (key-value, tabular, etc.) among visually-rich documents
Document AI Tasks

**Form Understanding**

| Key   | Value                     |
|-------|---------------------------|
| TO    | Lorillard Corporation    |
| ADDRESS | 666 Fifth Avenue      |
| CITY  | New York                 |
| Date  |                          |

**Receipt Understanding**

| Key         | Value                     |
|-------------|---------------------------|
| Total       | 4.95                      |
| Company     | Starbucks Store           |
| Address     | 11302 Euclid Avenue       |
| City        | Cleveland, OH             |
| Date        | 12/07/2014                |

**Document Image Classification**

Category: Form
Document AI Products

**Applications**
- Key-value Extraction
- Document Classification
- Document VQA

**Downstream Tasks**
- Table Detection
- Page Object Detection
- Reading Order Detection

**Benchmarks**
- TableBank (LREC’20)
- DocBank (COLING’20)
- ReadingBank (EMNLP’21)
- XFUND Benchmark

**Foundation Models**
- LayoutLM/LayoutLMv2/LayoutXLM
  (KDD’20, ACL’21)

https://aka.ms/document-ai/
LayoutLM
Model Family
LayoutLM -> LayoutLMv2 -> LayoutXLM

Text + Layout → Multi-modal → Multi-modal + Multi-lingual
LayoutLM: Pre-training of Text and Layout for Document Image Understanding, KDD'2020,
SPORTS MARKETING ENTERPRISES
DOCUMENT CLEARANCE SHEET

Date Routed: January 11, 1994
Contract No. 4011 00 00

Contract Subject: Joe's Place Exhibits
Company: SPEVCO, INC.
Brand(s): Camel/Winston
Total Contract Cost: $1,340,000.00
Current Year Cost: 1994-1995
Brief Description: Joe's Place Exhibits for use at Winston Cup, Winston Drag and Camel Super Bike Events.

G/T Code: Program Budget Code

Originator: Michael Wright
Manager: John Powell

REVIEW ROUTING

Insurance
Law
FS - Marketing

REVISIONS TO SHELL
(Other than Tarm., Compensation or Job)

APPROVAL ROUTING
* Sr. Manager (B. J. Powell)
* Director - (G. L. Littell)
** Sr. VP T. W. Robertson

Return To: MARY SEAGRAVES Ext. 1485

* UP TO AND INCLUDING $25,000
** OVER $25,000

Revised 10/26/92

Date Routed: January 11, 1994
Contract No. 4011 00 00

OCR/PDF Parser

| Image          | Token | Bounding Box (x0,y0,x1,y1) |
|----------------|-------|-----------------------------|
| Date           | Date  | 86 138 112 148              |
| Routed:        | Routed: | 117 138 162 148             |
| January        | January | 227 138 277 153             |
| 11             | 11    | 281 138 293 148             |
| 1994           | 1994  | 303 139 331 149             |
| Contract       | Contract | 415 138 464 149             |
| No.            | No.   | 468 139 487 149             |
| 4011           | 4011  | 556 139 583 150             |
| 00 00          | 0000  | 589 139 621 150             |

Input

Tok 0
Tok 1
Tok 2
Tok 3
Tok 4
Tok 5
Tok 6
Tok 7

Token Embeddings

Position Embeddings (x0)

Position Embeddings (y0)

Position Embeddings (x1)

Position Embeddings (y1)
Pre-training Data

11 million scanned document images from IIT-CDIP Test Collection 1.0
https://ir.nist.gov/cdip/
LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding, ACL'2021
Pre-training Tasks

• Masked Visual-Language Modeling
• Text-Image Matching
• Text-Image Alignment
LayoutLMv2 -> LayoutXLM

Multi-modal

Multi-modal + Multi-lingual
LayoutXLM: Multi-modal Pre-training for Multi-lingual Document Understanding, Preprint
Language Distribution for Pre-training

Totally 30M document images with 50+ languages used for pre-training the LayoutXLM
Document parsing meets 😊 Transformers!

#LayoutLMv2 and #LayoutXLM by @MSFTResearch are now available! 🔥

They're capable of parsing document images (like PDFs) by incorporating text, layout, and visual information, as in the @gradio demo below ↓

huggingface.co/spaces/nielsr/...
Benchmark Datasets
TableBank: A Benchmark Dataset for Table Detection and Recognition, LREC’2020
Download: https://aka.ms/tablebank
LayoutReader: Pre-training of Text and Layout for Reading Order Detection, EMNLP’2021
Download: https://aka.ms/readingbank
XFUND

LayoutXLM: Multi-modal Pre-training for Multi-lingual Document Understanding, Preprint
Download: https://aka.ms/xfund
Applications
Applications

- **Information Extraction**
  - Form Understanding (*FUNSD*)
    - [https://guillaumejaume.github.io/FUNSD/](https://guillaumejaume.github.io/FUNSD/)
  - Receipt Understanding (*SROIE, CORD*)
    - [https://rrc.cvc.uab.es/?ch=13](https://rrc.cvc.uab.es/?ch=13)
    - [https://github.com/clovaai/cord](https://github.com/clovaai/cord)
  - Document Information Extraction (*Kleister-NDA*)
    - [https://github.com/applicaai/kleister-nda](https://github.com/applicaai/kleister-nda)
  - Multi-lingual Form Understanding (*XFUND*)
    - [https://aka.ms/xfund](https://aka.ms/xfund)

- **Classification**
  - Document Image Classification (*RVL-CDIP*)
    - [https://www.cs.cmu.edu/~aharley/rvl-cdip/](https://www.cs.cmu.edu/~aharley/rvl-cdip/)

- **VQA**
  - Document Visual Question Answering (*DocVQA*)
    - [https://rrc.cvc.uab.es/?ch=17](https://rrc.cvc.uab.es/?ch=17)

- **Layout Analysis**
  - Table Detection (*TableBank*)
    - [https://aka.ms/tablebank](https://aka.ms/tablebank)
  - Page Object Detection (*DocBank*)
    - [https://aka.ms/docbank](https://aka.ms/docbank)
  - Reading Order Detection (*ReadingBank*)
    - [https://aka.ms/readingbank](https://aka.ms/readingbank)
| Model               | FUNSD | CORD  | SROIE | Kleister-NDA |
|---------------------|-------|-------|-------|--------------|
| BERT\_BASE         | 0.6026| 0.8968| 0.9099| 0.7790       |
| UniLM\_v2\_BASE    | 0.6648| 0.9092| 0.9459| 0.7950       |
| BERT\_LARGE        | 0.6563| 0.9025| 0.9200| 0.7910       |
| UniLM\_v2\_LARGE   | 0.7072| 0.9205| 0.9488| 0.8180       |
| LayoutLM\_BASE     | 0.7866| 0.9472| 0.9438| 0.8270       |
| LayoutLM\_LARGE    | 0.7895| 0.9493| 0.9524| 0.8340       |
| LayoutLM\_v2\_BASE | 0.8276| 0.9495| 0.9625| 0.8330       |
| LayoutLM\_v2\_LARGE| **0.8420**| **0.9601**| **0.9781**| **0.8520** |

Table 2: Entity-level F1 scores of the four entity extraction tasks: FUNSD, CORD, SROIE and Kleister-NDA.
| Model                          | Accuracy | #Parameters |
|-------------------------------|----------|-------------|
| BERT$_{\text{BASE}}$         | 89.81%   | 110M        |
| UniLMv2$_{\text{BASE}}$       | 90.06%   | 125M        |
| BERT$_{\text{LARGE}}$         | 89.92%   | 340M        |
| UniLMv2$_{\text{LARGE}}$      | 90.20%   | 355M        |
| LayoutLM$_{\text{BASE}}$ (w/ image) | 94.42% | 160M        |
| LayoutLM$_{\text{LARGE}}$ (w/ image) | 94.43% | 390M        |
| LayoutLMv2$_{\text{BASE}}$    | 95.25%   | 200M        |
| LayoutLMv2$_{\text{LARGE}}$   | 95.64%   | 426M        |
| VGG-16 (Afzal et al., 2017)   | 90.97%   | -           |
| Single model (Das et al., 2018)| 91.11% | -           |
| Ensemble (Das et al., 2018)   | 92.21%   | -           |
| InceptionResNetV2$^6$ (Szegedy et al., 2016) | 92.63% | -           |
| LadderNet (Sarkhel & Nandi, 2019) | 92.77% | -           |
| Single model (Dauphinee et al., 2019) | 93.03% | -           |
| Ensemble (Dauphinee et al., 2019) | 93.07% | -           |

Table 5: Classification accuracy on the RVL-CDIP dataset
| Model                  | Fine-tuning set | ANLS   | #Parameters |
|------------------------|----------------|--------|-------------|
| BERT\_BASE             | train          | 0.6354 | 110M        |
| UniLM\_v2\_BASE        | train          | 0.7134 | 125M        |
| BERT\_LARGE            | train          | 0.6768 | 340M        |
| UniLM\_v2\_LARGE       | train          | 0.7709 | 355M        |
| LayoutLM\_BASE         | train          | 0.6979 | 113M        |
| LayoutLM\_LARGE        | train          | 0.7259 | 343M        |
| LayoutLM\_v2\_BASE     | train          | 0.7808 | 200M        |
| LayoutLM\_v2\_LARGE    | train          | 0.8348 | 426M        |
| LayoutLM\_v2\_LARGE + QG | train + dev   | 0.8529 | 426M        |
| Top-1 on DocVQA Leaderboard (30 models ensemble)\(^7\) | -              | 0.8506 | -           |

Table 6: Average Normalized Levenshtein Similarity (ANLS) score on the DocVQA dataset (until 2020-12-24), “QG” denotes the data augmentation with the question generation dataset.
| Models                  | Word |          |          | Latex |          |          | Word+Latex |          |          |          |
|------------------------|------|----------|----------|-------|----------|----------|------------|----------|----------|----------|
|                        |      | Precision | Recall   | F1    | Precision| Recall   | F1         | Precision| Recall   | F1       |
| ResNeXt-101 (Word)     | 0.9496 | 0.8388   | 0.8908   | 0.9902 | 0.5948   | 0.7432   | 0.9594     | 0.7607   | 0.8486   |
| ResNeXt-152 (Word)     | 0.9530 | 0.8289   | 0.9166   | 0.9808 | 0.6890   | 0.8094   | 0.9603     | 0.8209   | 0.8851   |
| ResNeXt-101 (Latex)    | 0.8288 | 0.9395   | 0.8807   | 0.9854 | 0.9760   | 0.9807   | 0.8744     | 0.9512   | 0.9112   |
| ResNeXt-152 (Latex)    | 0.8259 | 0.9562   | 0.8863   | 0.9867 | 0.9754   | 0.9810   | 0.8720     | 0.9624   | 0.9149   |
| ResNeXt-101 (Word+Latex)| 0.9557 | 0.8403   | 0.8943   | 0.9886 | 0.9694   | 0.9789   | 0.9670     | 0.8817   | 0.9224   |
| ResNeXt-152 (Word+Latex)| 0.9540 | 0.8639   | 0.9067   | 0.9885 | 0.9732   | 0.9808   | 0.9657     | 0.8989   | 0.9311   |

Table 2: Evaluation results on Word and Latex datasets with ResNeXt-\{101,152\} as the backbone networks

| Models            | Precision | Recall | F1    |
|-------------------|------------|--------|-------|
| ICDAR 2013 (train)| 0.9748     | 0.7997 | 0.8786|
| UNLV              | 0.9185     | 0.9639 | 0.9406|
| Marmot            | 0.7692     | 0.9844 | 0.8636|
| DeepFigures       | 0.8527     | 0.9348 | 0.8918|
| TableBank (ResNeXt-152, Word) | 0.9725 | 0.8528 | 0.9087|
| TableBank (ResNeXt-152, Latex) | 0.9658 | 0.9594 | 0.9625|
| TableBank (ResNeXt-152, Word + Latex) | 0.9635 | 0.9039 | 0.9328|
| Tesseract         | 0.9439     | 0.7144 | 0.8133|
| Camelot           | 0.9785     | 0.6856 | 0.8063|

Table 3: Evaluation results on ICDAR 2013 dataset
Page Object Detection

| Models                  | Abstract | Author | Caption | Equation | Figure | Footer | List | Paragraph | Reference | Section | Table | Title | Macro average |
|-------------------------|----------|--------|---------|----------|--------|--------|------|-----------|-----------|---------|-------|-------|---------------|
| BERT_BASE               | 0.9394   | 0.8484 | 0.8629  | 0.8152   | 1.0000 | 0.7805 | 0.7133| 0.9619    | 0.9310    | 0.9081  | 0.8296 | 0.9442 | 0.8770        |
| RoBERTa_BASE            | 0.9288   | 0.8618 | 0.8944  | 0.8248   | 1.0000 | 0.8014 | 0.7353| 0.9646    | 0.9341    | 0.9337  | 0.8389 | 0.9511 | 0.8891        |
| LayoutLM_BASE           | **0.9816**| 0.8595 | 0.9597  | 0.8947   | 1.0000 | 0.8957 | 0.8948| 0.9738    | 0.9338    | 0.9598  | 0.8633 | **0.9979**    |
| BERT_LARGE              | 0.9286   | 0.8577 | 0.9650  | 0.8177   | 1.0000 | 0.7814 | 0.6960| 0.9619    | 0.9284    | 0.9065  | 0.8320 | 0.9430 | 0.8765        |
| RoBERTa_LARGE           | 0.9479   | 0.8724 | 0.9081  | 0.8370   | 1.0000 | 0.8392 | 0.7451| 0.9665    | 0.9334    | 0.9407  | 0.8494 | 0.9461 | 0.8988        |
| LayoutLM_LARGE          | 0.9784   | 0.8783 | 0.9556  | 0.8974   | 1.0000 | 0.9146 | 0.9004| 0.9790    | 0.9332    | 0.9596  | 0.8679 | 0.9552 | 0.9350        |

| X101                    | 0.9717   | 0.8227 | 0.9435  | 0.8938   | 0.8812 | 0.9029 | 0.9051| 0.9682    | 0.8798    | 0.9412  | 0.8353 | 0.9158 | 0.9051        |
| X101+LayoutLM_BASE      | 0.9815   | 0.8907 | **0.9669**| 0.9430   | 0.9990 | 0.9292 | **0.9300**| 0.9843    | **0.9437**| 0.8664  | 0.8818 | 0.9575 | 0.9478        |
| X101+LayoutLM_LARGE     | 0.9802   | 0.8964 | 0.9666  | **0.9440**| 0.9994 | **0.9352**| 0.9293| **0.9844**| **0.9430**| **0.9670**| **0.8875**| 0.9531 | **0.9488**    |

Table 4: The performance of BERT, RoBERTa, LayoutLM and Faster R-CNN on the DocBank test set.
| Method                     | Encoder   | Avg. Page-level BLEU ↑ | ARD ↓ |
|---------------------------|-----------|------------------------|-------|
| Heuristic Method          | -         | 0.6972                 | 8.46  |
| LayoutReader (text only)  | BERT      | 0.8510                 | 12.08 |
|                           | UniLM     | 0.8765                 | 10.65 |
| LayoutReader (layout only)| LayoutLM (layout only) | 0.9732               | 2.31  |
| LayoutReader              | LayoutLM  | 0.9819                 | 1.75  |

Table 2: Evaluation results of the LayoutReader on the reading order detection task, where the source-side of training/testing data is in the left-to-right and top-to-bottom order.

| Method             | Avg. Page-level BLEU ↑ | ARD ↓ |
|--------------------|------------------------|-------|
| Heuristic Method   | 0.3391                 | 13.61 |
| Tesseract OCR      | 0.7532                 | 1.42  |
| LayoutReader       | 0.9360                 | 0.27  |

Table 5: Adaptation to text lines of Tesseract OCR

| Method             | Avg. Page-level BLEU ↑ | ARD ↓ |
|--------------------|------------------------|-------|
| Heuristic Method   | 0.3752                 | 10.17 |
| The commercial OCR | 0.8530                 | 2.40  |
| LayoutReader       | 0.9430                 | 0.59  |

Table 6: Adaptation to text lines of the commercial OCR
Multilingual Document AI
### Table 2: Language-specific fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on X, testing on X), where “SER” denotes the semantic entity recognition and “RE” denotes the relation extraction.

**SER**: Semantic Entity Recognition (headers, keys, values)  
**RE**: Relation extraction for key-value pairs
|        | Model             | FUNSD | ZH   | JA   | ES   | FR   | IT   | DE   | PT   | Avg. |
|--------|-------------------|-------|------|------|------|------|------|------|------|------|
| **SER** | XLM-RoBERTa\_BASE | 0.667 | 0.4144 | 0.3023 | 0.3055 | 0.371 | 0.2767 | 0.3286 | 0.3936 | 0.3824 |
|        | InfoXLM\_BASE     | 0.6852 | 0.4408 | 0.3603 | 0.3102 | 0.4021 | 0.2880 | 0.3587 | 0.4502 | 0.4119 |
|        | LayoutXLM\_BASE   | **0.794** | **0.6019** | **0.4715** | **0.4565** | **0.5757** | **0.4846** | **0.5252** | **0.539** | **0.5561** |
| **RE**  | XLM-RoBERTa\_LARGE | 0.7074 | 0.5205 | 0.3939 | 0.3627 | 0.4672 | 0.3398 | 0.418 | 0.4997 | 0.4637 |
|        | InfoXLM\_LARGE    | 0.7325 | 0.5536 | 0.4132 | 0.3689 | 0.4909 | 0.3598 | 0.4363 | 0.5126 | 0.4835 |
|        | LayoutXLM\_LARGE  | **0.8225** | **0.6896** | **0.519** | **0.4976** | **0.6135** | **0.5517** | **0.5905** | **0.6077** | **0.6115** |

Table 3: Zero-shot transfer accuracy (F1) on the XFUND dataset (fine-tuning on FUNSD, testing on X), where “SER” denotes the semantic entity recognition and “RE” denotes the relation extraction.
### Multitask Learning

| Model               | FUNSD | ZH  | JA  | ES  | FR  | IT  | DE  | PT  | Avg. |
|---------------------|-------|-----|-----|-----|-----|-----|-----|-----|------|
| **SER**             |       |     |     |     |     |     |     |     |      |
| XLM-RoBERTaBASE     | 0.6633| 0.883| 0.7786| 0.6223| 0.7035| 0.6814| 0.7146| 0.6726| 0.7149|
| InfoXLM_BASE        | 0.6538| 0.8741| 0.7855| 0.5979| 0.7057| 0.6826| 0.7055| 0.6796| 0.7106|
| LayoutXLM_BASE      | **0.7924**| **0.8973**| **0.7964**| **0.7798**| **0.8173**| **0.821**| **0.8322**| **0.8241**| **0.8201**|
| XLM-RoBERTa_LARGE   | 0.7151| 0.8967| 0.7828| 0.6615| 0.7407| 0.7165| 0.7431| 0.7449| 0.7502|
| InfoXLM_LARGE       | 0.7246| 0.8919| 0.7998| 0.6702| 0.7376| 0.7180| 0.7523| 0.7332| 0.7534|
| LayoutXLM_LARGE     | **0.8068**| **0.9155**| **0.8216**| **0.8055**| **0.8384**| **0.8372**| **0.853**| **0.8650**| **0.8429**|
| **RE**              |       |     |     |     |     |     |     |     |      |
| XLM-RoBERTaBASE     | 0.3638| 0.6797| 0.6829| 0.6828| 0.6727| 0.6937| 0.6887| 0.6082| 0.6341|
| InfoXLM_BASE        | 0.3699| 0.6493| 0.6473| 0.6828| 0.6831| 0.6690| 0.6384| 0.5763| 0.6145|
| LayoutXLM_BASE      | **0.6671**| **0.8241**| **0.8142**| **0.8104**| **0.8221**| **0.8310**| **0.7854**| **0.7044**| **0.7823**|
| XLM-RoBERTa_LARGE   | 0.4246| 0.7316| 0.7350| 0.7513| 0.7532| 0.7520| 0.7111| 0.6582| 0.6896|
| InfoXLM_LARGE       | 0.4543| 0.7311| 0.7510| 0.7644| 0.7549| 0.7504| 0.7356| 0.6875| 0.7037|
| LayoutXLM_LARGE     | **0.7683**| **0.9000**| **0.8621**| **0.8592**| **0.8669**| **0.8675**| **0.8263**| **0.8160**| **0.8458**|

Table 4: Multitask fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on 8 languages all, testing on X), where “SER” denotes the semantic entity recognition and “RE” denotes the relation extraction.
Challenges in Document AI

- Model Limitations
- Data Quality in Real-Word Tasks
- Task Correlations
- Data/Computation Insufficiency
Document AI @MSRA

• Multimodal Pre-trained Models
  • \textit{LayoutLM} (KDD’2020)
  • \textit{LayoutLMv2} (ACL’2021)
  • \textit{LayoutXLM} (Preprint)

• Benchmark Datasets
  • \textit{TableBank} (LREC’2020)
  • \textit{DocBank} (COLING’2020)
  • \textit{ReadingBank} (EMNLP’2021)
  • \textit{XFUND} (with LayoutXLM)

• Our paper \textit{“Document AI: Benchmarks, Models and Applications”} will be publicly available soon