Document AI: Benchmarks, Models, and Applications

Summary:

This paper discusses a comprehensive survey of Document AI (Document Intelligence), discussing its evolution, benchmarks, and emerging trends. Document AI refers to automatic document reading, understanding, and analysis using deep learning and natural language processing (NLP).

The paper explores:

Traditional methods: Early rule-based and statistical machine learning models.

Deep learning-based approaches: Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and Transformers.

Pre-training methods: Multi-modal transformers like LayoutLM and LayoutLMv2 for visually-rich document understanding.

Implementation Methodology:

The paper discusses various Document AI tasks, including:

Document Layout Analysis – Detecting tables, images, and text structures using Faster R-CNN, SSD, and YOLO.

Visual Information Extraction – Extracting key entities using Graph Neural Networks (GNNs).

Document Visual Question Answering (DocVQA) – Answering questions based on document images.

Document Image Classification – Categorizing documents into different types.

Key models discussed:

CNN-based models (e.g., Faster R-CNN for layout detection).

GNN-based models (e.g., Graph Convolution Networks for relation extraction).

Transformer-based models (e.g., LayoutLM, which integrates text, layout, and vision).

DocFormer: End-to-End Transformer for Document Understanding

Summary

This paper Introduces DocFormer, a multi-modal transformer model designed for Visual Document Understanding (VDU). The model combines text, vision, and spatial features using a novel multi-modal self-attention mechanism that enables better cross-modal feature fusion. Unlike previous approaches that rely on pre-trained object detection models, DocFormer uses ResNet50 for feature extraction and trains the entire pipeline end-to-end, making it more efficient.
The Paper also proposes 3 new unsupervised pre-training task:

1. **Multi-Modal Masked Language Modeling (MM-MLM):** Encourages interaction between text and vision.
2. **Learn-to-Reconstruct (LTR):** Uses multi-modal representations to reconstruct the original image.
3. **Text Describes Image (TDI):** A binary classification task that determines if the provided text corresponds to the given image.

Implementation Methodology:

The architecture consists of a transformer encoder with 12 layers, integrating textual, spatial, and visual embeddings. ResNet50 extracts image features, and LayoutLMv1 embeddings initialize the text components. Shared spatial embeddings help connect text and image features, improving alignment.

Pre-training: The model is trained on a subset of 5 million pages from the IIT-CDIP dataset.

Fine-tuning: DocFormer is evaluated on four datasets—FUNSD, CORD, Kleister-NDA, and RVL-CDIP.

6.1. Implementation Details

We present all the hyper-parameters in Table 10 used for pre-training and fine-tuning DocFormer . We fine-tune on downstream tasks on the same number of epochs as prior art [56, 57, 26]: FUNSD [18], Kleister-NDA [17] datasets were fine-tuned for 100 epochs. CORD [47] for 200 epochs. RVL-CDIP [20] for 30 epochs. For Key, Query 1-D relative local attention we choose a span of 8 i.e. for a particular multi-modal feature, DocFormer gives more attention 8 tokens to its left and right.

Hyper-Parameter	Pre-training	Fine-tuning
Epochs	5	varies
Learning rate	5E-05	2.5E-05
Warm-up	10% iters	0
Gradient Clipping	1.0	1.0
Gradient agg.	False	False
Optimizer	AdamW[37]	AdamW[37]
Lower case	True	True
Sequence length	512	512
Encoder layers	12	12
32-bit mixed precision	True	True
Batch size	9 per GPU	4 per GPU
GPU hardware	A100 (40GB)	V100 (16GB)
Training Num. Samples	5M	varies
Training time	17 hours/epoch	varies

Table 10: **Implementation Details**: Hyper-parameters used for pre-training DocFormer and fine-tuning for downstream tasks. Training epochs vary for down-stream tasks.

Evaluating Deep Neural Networks for Image Document Enhancement

Summary:

This paper evaluates various deep Convolutional Neural Networks (CNNs) for document image classification, comparing their performance on RVL-CDIP and Tobacco-3482 datasets. It also investigates transfer learning, comparing models pre-trained on ImageNet vs. document datasets.

Implementation Methodology:

Datasets Used:

RVL-CDIP: 400,000 document images from 16 classes.

Tobacco-3482: 3,482 document images from 10 classes.

CNN Architectures Evaluated: AlexNet, GoogLeNet, VGG-16, ResNet-50

Training Strategies:

Models were trained using stochastic gradient descent with momentum.

Pre-training on document datasets vs. ImageNet pretraining was compared.

Experiments were conducted with different training sample sizes.

Results:

VGG-16 performed best on RVL-CDIP, achieving 90.97% accuracy, outperforming prior state-of-the-art models.

ResNet-50 performed best on Tobacco-3482, reaching 91.13% accuracy when pre-trained on document images.

Pre-training on document datasets significantly improved classification accuracy.

Transfer learning helped even with small datasets, proving document-specific features are crucial for accurate classification.

LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding

Summary:

This paper proposes, LiLT is a language-independent transformer model designed for structured document understanding (SDU). Unlike previous models that require pre-training on multiple languages, LiLT can pre-train on a single language (English) and generalize to others without additional pre-training. It achieves this by decoupling text and layout information, optimizing them separately, and then re-integrating them during fine-tuning.

Implementation Methodology:

Model Architecture:

Dual-stream Transformer: One for text representation (using RoBERTa, XLM-R, InfoXLM) and another for layout representation.

Introduces Bi-directional Attention Complementation Mechanism (BiACM) to enhance cross-modal learning.

Pre-training Tasks:

Masked Visual-Language Modeling (MVLM): Predicts missing words using layout-aware context.

Key Point Location (KPL): Identifies the spatial placement of text elements in documents.

Cross-modal Alignment Identification (CAI): Detects whether textual and spatial information align correctly.

Fine-tuning: LiLT is adapted to multiple languages by combining it with pre-trained text models.

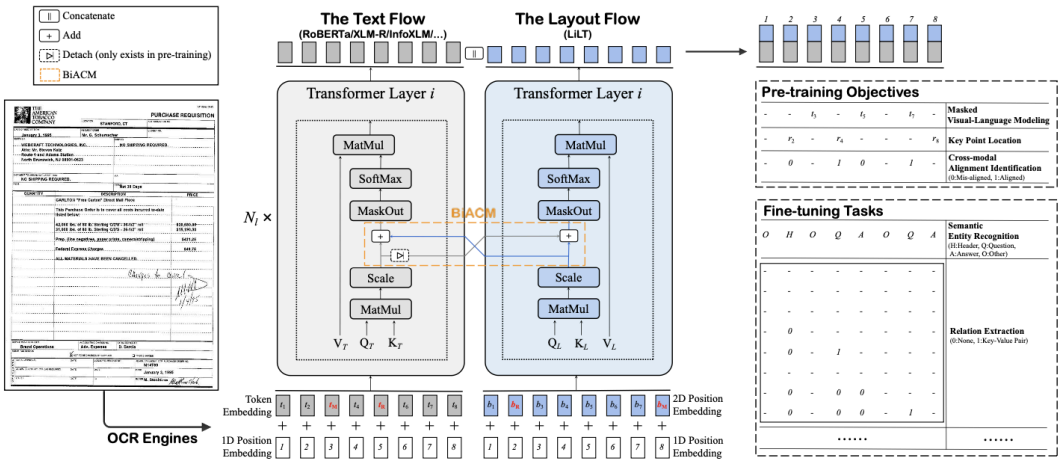

Figure 2: The overall illustration of our framework. Text and layout information are separately embedded and fed into the corresponding flow. BiACM is proposed to accomplish the cross-modality interaction. At the model output, text and layout features are concatenated for the self-supervised pre-training or the downstream fine-tuning. N_l is the number of Transformer layers. The red $*_{\mathrm{M}}/*_{\mathrm{R}}$ indicates the randomly masked/replaced item for pre-training. t, b and r represent $token$, box and $region$, respectively. Best viewed in zoomed-in.

Multi-document Summarization via Deep Learning Techniques: A Survey

Summary:

This paper provides a comprehensive survey of multi-document summarization (MDS) using deep learning techniques. MDS aims to generate a concise and informative summary from a cluster of related documents, which is more complex than single-document summarization (SDS) due to redundancy, contradiction, and diverse perspectives.

Survey:

Explored deep learning architectures for MDS, including:

Recurrent Neural Networks (RNNs): Used in early models but struggled with long dependencies.

Convolutional Neural Networks (CNNs): Extracted semantic-rich representations.

Graph Neural Networks (GNNs): Modeled cross-document relationships effectively.

Pre-trained Transformer Models (BERT, GPT, BART, PEGASUS): Achieved state-of-the-art results in recent years.

Evaluation Metrics Discussed: ROUGE, BLEU, METEOR, BERTScore.

Results:

Transformer-based models significantly outperform traditional methods in MDS. Graph-based methods (GCNs, GATs) improve coherence by modeling relationships between sentences.

dhSegment: A Generic Deep-Learning Approach for Document Segmentation

Summary:

This paper proposes, dhSegment which is a flexible deep-learning-based document segmentation model that can handle various tasks, including page extraction, layout analysis, baseline detection, and illustration extraction. Unlike task-specific methods, dhSegment is generic, allowing a single CNN architecture to perform multiple document segmentation tasks with minimal post-processing.

Implementation Methodology:

Model Architecture:

Based on a Fully Convolutional Network (FCN) with a ResNet-50 encoder and a U-Net-inspired decoder.

Outputs pixel-wise segmentation masks, which are post-processed for different tasks.

Training Process:

Uses pre-trained weights from ResNet-50 for feature extraction.

Applies data augmentation (rotation, scaling, mirroring) for better generalization.

Post-processing techniques: Thresholding, morphological operations, and connected component analysis refine segmentation results.

Evaluation Tasks:

Page Extraction: Detects document boundaries.

Baseline Detection: Identifies handwritten text lines.

Illustration Extraction: Segments illustrations from historical documents.

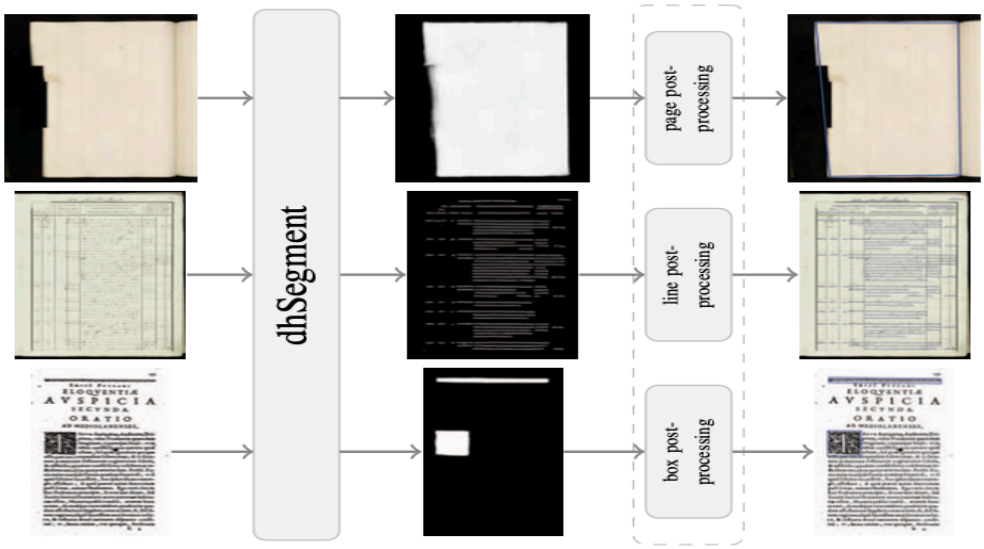

Fig. 1. Overview of the system. From an input image, the generic neural network (dhSegment) outputs probabilities maps, which are then post-processed to obtain the desired output for each task.

Results:

Page Extraction: Achieved 98% accuracy, outperforming previous models.

Baseline Detection: Showed superior performance over traditional methods.

Generalization Ability: Successfully applied to various historical document datasets with minor adjustments.