Deep Learning for Document Understanding: A Comprehensive Survey

Abstract

This paper presents a comprehensive survey of deep learning techniques applied to document understanding. We review recent advances in neural network architectures, training methodologies, and evaluation frameworks that have significantly improved the state-of-the-art in document analysis tasks.

1. Introduction

Document understanding has emerged as a critical task in artificial intelligence, with applications ranging from automated information extraction to intelligent document processing systems. Traditional approaches relied heavily on handcrafted features and rule-based systems, which often struggled with the complexity and variability of real-world documents.

1.1 Problem Definition

Document understanding encompasses several interconnected tasks including text extraction, layout analysis, semantic segmentation, and content classification. Each of these tasks presents unique challenges related to document diversity and structural complexity.

1.2 Scope and Contributions

This survey focuses on deep learning approaches developed in the past five years. Our main contributions include: (1) a comprehensive taxonomy of neural architectures, (2) analysis of training strategies, and (3) evaluation of current benchmarks and datasets.

2. Related Work

Early work in document understanding relied on traditional computer vision and natural language processing techniques. Optical Character Recognition (OCR) systems formed the foundation for text extraction, while geometric analysis methods were used for layout understanding.

2.1 Traditional Approaches

Traditional document analysis systems typically followed a pipeline approach: preprocessing, segmentation, feature extraction, and classification. While effective for structured documents, these methods often failed on complex layouts and noisy inputs.

2.1.1 OCR Systems

Optical Character Recognition systems have evolved from simple template matching to sophisticated machine learning approaches. Modern OCR engines incorporate neural networks for improved accuracy on diverse fonts and imaging conditions.

2.2 Deep Learning Revolution

The introduction of deep learning transformed document understanding by enabling end-to-end learning of complex patterns and relationships. Convolutional neural networks proved particularly effective for visual document analysis tasks.

3. Methodology

Our approach to document understanding combines multiple neural network architectures in a unified framework. We leverage both visual and textual information through multi-modal learning techniques.

3.1 Neural Architecture Design

We propose a hierarchical neural architecture that processes documents at multiple granularities: pixel-level for visual features, word-level for semantic understanding, and document-level for global structure comprehension.

3.1.1 Visual Feature Extraction

The visual component employs a ResNet-based backbone for extracting visual features from document images. Feature pyramid networks are used to capture multi-scale information essential for handling documents with varying text sizes and layout complexity.

3.1.2 Text Processing Module

Text features are processed using transformer-based encoders that capture long-range dependencies within documents. Positional encodings incorporate spatial information to maintain layout awareness.

3.2 Training Strategy

We employ a multi-task learning framework that jointly optimizes for text detection, recognition, and layout analysis. This approach enables the model to learn shared representations that benefit all downstream tasks.

4. Experiments

We evaluate our approach on standard document understanding benchmarks including DocVQA, FUNSD, and CORD datasets. Experimental results demonstrate significant improvements over existing methods.

4.1 Datasets and Metrics

Our evaluation covers diverse document types including forms, receipts, research papers, and administrative documents. Performance is measured using standard metrics including F1-score, exact match accuracy, and BLEU scores for text generation tasks.

4.2 Comparative Analysis

Comparison with state-of-the-art methods shows consistent improvements across all evaluation metrics. Our approach achieves particularly strong performance on complex document layouts that challenge traditional pipeline-based systems.

5. Conclusion

This work presents a comprehensive framework for document understanding that leverages the latest advances in deep learning. Our multi-modal approach achieves state-of-the-art results while maintaining computational efficiency suitable for practical applications.

5.1 Future Work

Future research directions include extending the framework to handle multilingual documents, incorporating external knowledge bases, and developing more efficient architectures for real-time processing applications.