

ABSTRACT

This paper presents a novel approach to automated document processing using machine learning techniques. We introduce a three-stage pipeline that achieves state-of-the-art performance on benchmark datasets.

1. INTRODUCTION

Document processing has become increasingly important in the digital age. Traditional methods rely on rule-based systems, but our approach leverages deep learning to extract meaningful information from complex documents. This work builds upon recent advances in transformer architectures.

2. METHODOLOGY

We propose a three-stage pipeline:

1. Text extraction using optical character recognition (OCR)
2. Structure analysis using convolutional neural networks
3. Content classification using transformer models

The methodology involves training a deep convolutional neural network on a dataset of 10,000 labeled documents. We use a ResNet-50 architecture with custom attention mechanisms for improved performance.

Our approach processes documents through the following algorithm:

- Algorithm 1: Document preprocessing
- Algorithm 2: Feature extraction
- Algorithm 3: Classification and ranking

3. EXPERIMENTAL RESULTS

Our approach achieved 95.2% accuracy on the test dataset, significantly outperforming baseline methods:

- Baseline SVM: 78.3%
- Random Forest: 82.1%
- BERT baseline: 91.4%
- Our method: 95.2%

The methodology was evaluated on three different document types: academic papers, technical reports, and legal documents.

4. DISCUSSION

The results demonstrate the effectiveness of our proposed methodology for automated document processing. The attention mechanism proves particularly useful for handling complex document layouts.

5. CONCLUSION

This work demonstrates the effectiveness of our proposed methodology for automated document processing. Future work will focus on extending the approach to multilingual documents and real-time processing.

REFERENCES

- [1] Smith, J. (2020). Document Analysis Techniques. Journal of AI Research, 15(3), 123-145.
- [2] Brown, A. et al. (2021). Transformer Models for Document Processing. ICML 2021.
- [3] Wilson, K. (2019). Deep Learning in Document Analysis. Nature Machine Intelligence.