Visual Document Matching for Zero-Shot Document Classification

Lucas De Almeida Bandeira Macedo ✉[1 [0009−0004−6952−4731]], Joao Paulo Vieira Costa ✉[1 [0000−0001−7379−0961]], Joao Pedro Felix De Almeida[1 [0009−0008−3367−8033]], Pedro Garcia Freitas[1 [0000−0003−0866−658X]], and Li Weigang[1 [0000−0003−1826−1850]]

University of Brasília, UnB - Brasília, Brasilia - Federal District, 70910-900

Abstract. Accurate document identification is crucial for ensuring compliance and maintaining consistency across various applications. Consequently, document classification has been extensively studied. However, the dynamic nature of documents often renders simple classification insufficient, necessitating frequent model retraining. To mitigate this constant maintenance, comparing incoming documents against known references is a more suitable approach. To address the zero-shot classification problem, we introduce Visual Document Matching (VDM). VDM focuses on verifying whether two documents share an identical visual layout structure, a particularly effective approach in scenarios where training classes do not align with those encountered during inference. Despite its significant potential, the Zero-Shot Learning (ZSL) approach remains largely underexplored in document layout understanding. To support our study, we introduce Layout-Aware Complex Document Information Processing (LA-CDIP), a dataset comprising 4,993 documents across 144 classes. We reorganized this dataset from the RVL-CDIP database to emphasize visual structure over semantic information. Our approach is benchmarked using a siamese network within a contrastive learning framework across multiple backbone architectures, including ResNet, EfficientNet, and Vision Transformer (ViT). In zero-shot scenarios, our method achieves an Equal Error Rate (EER) below 5% in 1-vs-1 verification with cross-validation. Furthermore, our VDM approach outperforms lighter Large Language Models (LLMs) and rivals GPT-4o. These findings highlight the superiority of specialized VDM techniques over general-purpose multimodal models, demonstrating high accuracy with significantly fewer parameters, making our approach more practical for real-world applications.

Keywords: Zero-shot learning · Metric Learning · Document Understanding · Document Image Classification · Visual Document Matching

1 Introduction

Documents are central to many organizations, acting as official records for contracts, identities, and evidence, and playing a crucial role in legal, financial, and