Best Papers Related to RVL-CDIP : aharley/rvl_cdip · Datasets at Hugging Face:

1. EAML: Ensemble Self-Attention-based Mutual Learning Network for Document Image Classification

   Summary: This research article introduces EAML, a novel framework designed to enhance multi-modal learning, particularly in scenarios where different modalities (e.g., text, image, audio) need to be integrated for better performance on tasks like classification and regression. The key innovation of EAML is its use of ensemble self-attention mechanisms that enable effective feature extraction by learning cross-modal relationships. The authors demonstrate that their approach facilitates better mutual learning among various modalities, leading to improved representation learning. This method operates by leveraging self-attention to weigh the contributions of different modalities dynamically, enhancing the model's capability to make informed predictions. The paper includes thorough experiments showcasing the effectiveness of EAML across various datasets, emphasizing its advantages over traditional multi-modal learning methods.

2. VisualWordGrid: Information Extraction From Scanned Documents Using A Multimodal Approach

   Summary: This paper presents VisualWordGrid, a novel method for information extraction from scanned documents that integrates textual, visual, and layout information into a unified 3-axis tensor representation. This approach enhances the performance of field extraction tasks, particularly for documents with complex visual elements such as tables and logos. By improving upon existing models like Chargrid and Wordgrid, VisualWordGrid demonstrates robustness even with smaller datasets while maintaining low inference times.
   The introduction of a multimodal representation that combines textual embeddings with visual data from document images. Two model architectures: VisualWordGrid-pad, which uses a padding strategy to incorporate visual information, and VisualWordGrid-2encoders, which employs two separate encoders for textual and visual data. Extensive experiments on both public (RVL-CDIP) and private (Tax Notice) datasets, showing that VisualWordGrid

outperforms traditional rule-based systems and other state-of-the-art methods in terms of accuracy and efficiency.

3. Modular Multimodal Architecture for Document Classification
   Summary: The paper presents a modular multimodal architecture for document classification that combines visual (image-based) and textual (OCR-extracted) content analysis. Using late fusion, it integrates predictions from a CNN image classifier and a text classifier (based on Bag-of-Words) through a meta-classifier (XGBoost). This architecture achieves 93.03% test accuracy on the RVL-CDIP dataset, surpassing state-of-the-art benchmarks. Its modularity facilitates flexible model replacement or addition without retraining other components.

Best Paper Related to : PubLayNet - IBM Developer

1. Vision Grid Transformer for Document Layout Analysis

   Summary: The paper introduces the Vision Grid Transformer (VGT), a two-stream multi-modal model designed for Document Layout Analysis (DLA). VGT integrates a Vision Transformer (ViT) for image-based features and a novel Grid Transformer (GiT) for text-based features. By leveraging pre-training tasks like Masked Grid Language Modeling (MGLM) and Segment Language Modeling (SLM), it captures both token-level and segment-level semantics. The proposed D4LA dataset, featuring diverse document types and layouts, further highlights VGT's ability to achieve state-of-the-art results on DLA benchmarks