Pattern Recognition, vol.172, no.Part B, pp.112502-1-112502-14, April, 2026.

The MERIT Dataset: Modelling and efficiently rendering interpretable transcripts

I. de Rodrigo Tobías; A. Sánchez Cuadrado; J. Boal Martín-Larrauri; A.J. López López

Abstract-

This paper introduces the MERIT Dataset, a multimodal, fully labeled dataset of school grade reports. Comprising over 400 labels and 33k samples, the MERIT Dataset is a resource for training models in demanding Visually-rich Document Understanding tasks. It contains multimodal features that link patterns in the textual, visual, and layout domains. The MERIT Dataset also includes biases in a controlled way, making it a valuable tool to benchmark biases induced in Language Models. The paper outlines the dataset’s generation pipeline and highlights its main features and patterns in its different domains. We benchmark the dataset for token classification, showing that it poses a significant challenge even for SOTA models.

Index Terms- Synthetic Dataset; Multimodal Dataset; Visually-rich Document Understanding; Vision-Language Models

Citation:
Rodrigo, I. de; Sánchez-Cuadrado, A.; Boal, J.; López López, A.J. "The MERIT Dataset: Modelling and efficiently rendering interpretable transcripts", Pattern Recognition, vol.172, no.Part B, pp.112502-1-112502-14, April, 2026.