CS230

A Neural Verifier for Structured Table Extraction

Ray Hu
rayhu@stanford.edu

Nofel Teldjoune
nofelt@stanford.edu

Hiva Zaad
hiva@stanfurd.edu

1 Problem Description

Large Language Models (LLMs) and vision-language models have achieved remarkable performance in document understanding, yet remain unstable in structured data extraction. For example, while models such as TATR-v1.0 achieve near-perfect structural recognition on PubTables-1M (GriTS ≈ 0.985), their semantic content accuracy remains imperfect ($\mathrm{Acc}_{\mathrm{Con}} \approx 0.82$).

Even state-of-the-art multimodal systems such as GPT-4o reach less than 90% Grid Table Similarity (GriTS F1) (1), leaving a nontrivial reliability gap.

This suggests that while table geometry is largely solved, semantic correctness and consistency across extracted cells still present challenges.

This instability limits the use of LLMs in high-stakes workflows such as legal, financial, or medical document analysis, where structured accuracy and auditability are essential.

The goal of this project is to train a neural **Verifier model** that automatically evaluates the quality of structured table extraction from PDFs. Instead of manually comparing extracted tables to ground truth, the model will learn to predict extraction quality—quantifying semantic correctness, schema alignment, and numerical consistency. This Verifier can serve as a foundation for improving reliability, benchmarking models, or guiding future optimization.

Input and Output Definition:

- **Input:** A raw PDF document and its corresponding extracted table in structured format (e.g., JSON or Markdown table).
- **Process:** The Verifier model receives both the extracted table and reference features (or derived statistics) and outputs a quality score based on multiple criteria such as schema validity, field completeness, and numeric consistency.
- **Output:** A scalar score $r \in [0, 1]$ representing the predicted extraction quality.

This approach reframes table extraction evaluation as a supervised learning problem, enabling scalable and automated quality assessment without human inspection.

2 Related Work

Recent advances in document AI have significantly improved multimodal understanding. LayoutLMv3 (2) and Donut (3) achieve strong OCR-free document parsing capabilities. Silva et al. (2024) benchmarked a wide range of models on PubTables-1M, showing that even the top-performing GPT-4o achieved 89.6% GriTS F1 (1). However, current models lack mechanisms for *automatic*

CS230: Deep Learning, Autumn 2025, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

evaluation or reliability estimation. Most systems rely on costly ground-truth comparisons, which do not generalize across datasets or domains.

Our approach introduces a learned evaluator—a neural Verifier—to predict extraction quality directly from outputs, bridging the gap between model accuracy and practical reliability assessment.

3 Dataset and Features

We will use the **PubTables-1M** dataset (4), a large-scale benchmark containing over one million annotated tables extracted from scientific and government PDFs. It provides ground-truth table structures, content, and metadata suitable for both training and evaluation of extraction and verification models.

A manageable subset (e.g., 10,000 tables) will be sampled for this project. Synthetic variations and weak supervision will be applied to generate a range of "good" and "imperfect" extraction examples for Verifier training.

Evaluation Metrics:

- **Verifier Accuracy:** Agreement between predicted and true correctness labels.
- **Field-level F1:** Comparison of predicted and ground-truth table content.
- **Calibration Quality:** How well predicted scores correlate with actual performance.

4 Proposed Method

- **1. Verifier Model:** We will train a lightweight neural network $V_\theta(x)$ that predicts a scalar quality score in [0,1] for each extracted table. Input features will include schema similarity, content consistency, and cell-level matching statistics, as well as optional embeddings of table text.
- **2. Data Generation:** Multiple extraction candidates will be generated using open-source models such as TATR, Donut, or LayoutLMv3. Ground-truth comparisons will yield weak labels for Verifier training, enabling supervision without manual annotation.
- **3. Evaluation:** The trained Verifier will be evaluated on held-out data to assess its ability to rank and score extractions accurately. We will analyze score calibration, AUC, and pairwise ranking accuracy to validate its reliability as an automated evaluator.

References

[1] Silva et al., *Benchmarking Table Extraction: A Comparative Study of Multimodal Models*, Universidade NOVA de Lisboa, 2024. `https://run.unl.pt/bitstream/10362/185720/1/Benchmarking_table_extraction.pdf`

[2] Xu et al., *LayoutLMv3: Pretraining for Document AI*, arXiv, 2023.

[3] Kim et al., *Donut: OCR-Free Document Understanding Transformer*, ECCV, 2022.

[4] Smock et al., *PubTables-1M: Towards Comprehensive Table Extraction*, CVPR, 2022.