Short Diploma Summary

Using the Fatura Dataset for Invoice Field Extraction

1. Introduction.

I conducted a research on *invoice field extraction* using the publicly available *Fatura* dataset [1], utilizing traditional ML solutions. This dataset includes invoice documents with labeled text and bounding boxes, enabling comprehensive experiments on multiple models and features.

2. Data Processing and Feature Engineering.

- *Bounding Box Parsing:* Leveraged bounding-box coordinates (`x_min`, `y_min`, etc.) and normalized them for spatial features.
- *Text Embeddings:* Used TF-IDF (`TfidfVectorizer`) to encode invoice text into numerical representations.
- *Combined Embeddings:* Merged text and normalized spatial features into a single feature matrix.

3. Model Training.

- *Classifiers:* Evaluated Logistic Regression, SVM (linear, RBF), and XGBoost.
- *Oversampling:* Addressed class imbalance with `RandomOverSampler` on minority fields (e.g., GST).
- *Validation Splits:* Reserved a portion of data (e.g., 400 invoices) to confirm generalization.

4. Evaluation Strategies.

- *Intra-Template-Centric:* Each of 50 templates is split into train/valid, all 50 templates used during training and validation.
- *Inter-Template-Centric:* Some templates held out entirely during training; models face unseen layouts during validation.

5. Results and Observations.

- *High Accuracy:* Models can achieve 100% accuracy and F1-scores on familiar invoice patterns and classes (*Intra-Template-Centric scenario*).
- *Comparison:* Both traditional (SVM, Logistic Regression) and ensemble methods (XGBoost) performed strongly.
- *Generalization:* Performance remains high on unseen but structurally similar data (intra-template). More diverse or completely new invoice layouts (inter-template) may require additional or more sophisticated techniques.

6. Conclusion.

Traditional machine learning solutions can achieve high accuracy on data that closely resembles the training set, making them an excellent choice for environments where known classes and patterns dominate. In such scenarios, training the model on representative samples before deployment leads to strong, reliable performance. However, these models can lack flexibility when encountering substantially different or unforeseen data. Consequently, for projects requiring adaptability to evolving layouts or novel classes, further exploration into deep learning approaches (e.g., Transformers or multimodal architectures) could offer more robust generalization and advanced text-layout synergy.

References

References

[1] Göker, Z., et al. (2023). *Fatura: A Large Real Invoice Dataset.* Available at `https://arxiv.org/pdf/2311.11856`.