LayerX

バクラクのドキュメント解析技術と実データにおける課題

2024/12/14

cvpaper.challenge Conference winter 2024

株式会社 LayerX　Naoto Shimakoshi（@nt_4o54）

島越 直人（Naoto Shimakoshi)

バクラク事業部 AI-OCRグループ Tech Lead/ 機械学習エンジニア

経歴

- 2019/04 京都大学大学院 工学研究科 修士課程修了
- 新卒では、DeNAでタクシー配車アプリに関する機械学習システムの構築や、ライブストリーミングサービスにおける推薦システム構築に携わる
- 現在
  - 株式会社LayerX AI-OCRグループ Tech Lead
  - バクラク事業部において、AI-OCRの改善や新しい機械学習システムの構築を担当
  - Kaggle Competitions Grandmaster

Agenda

目次

- LayerXについて
- Introduction
- Bakuraku AIの紹介
- 実データにおける課題
- 今後の展望

LayerXについて

**「すべての経済活動を、デジタル化する。」をミッションに掲げ、
法人支出管理サービス「バクラク」や企業内業務のデジタル化を支援するサービスを提供しています。**

バクラク事業

企業活動のインフラとなる法人支出管理（BSM）SaaSを開発・提供

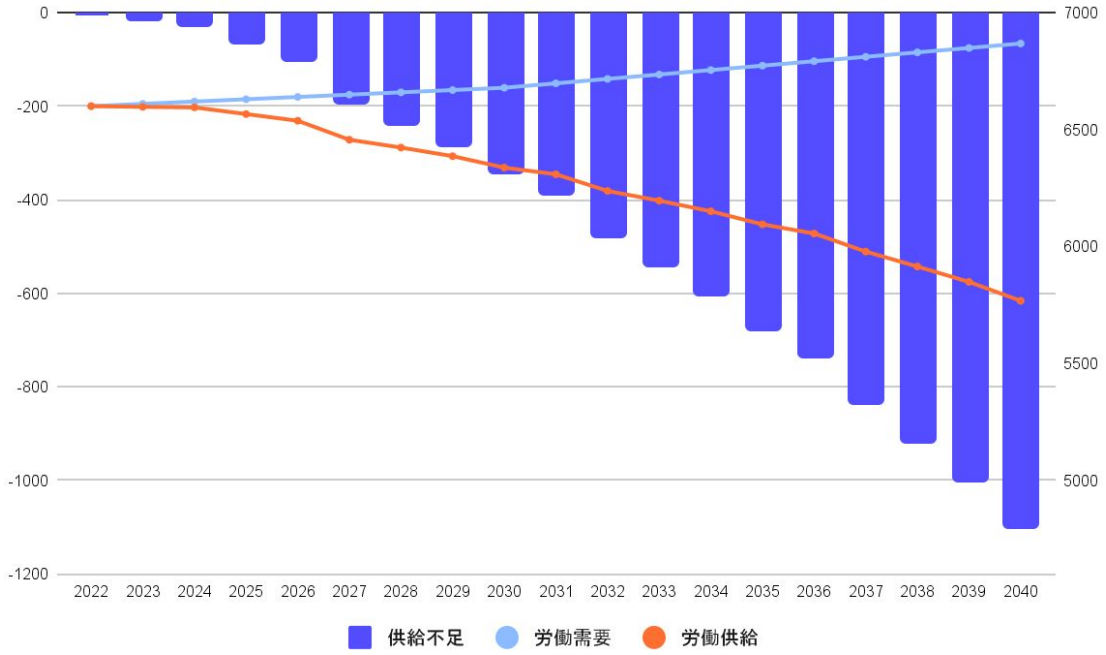
Fintech事業

ソフトウェアを駆使したアセットマネジメント・証券事業を合弁会社にて展開

AI・LLM事業

文書処理を中心とした、LLMの活用によるプロセスのリデザイン

日本の労働需給ギャップは深刻



2040年に労働需給ギャップ

1100万人

日本全体で必要な生産性

+20%

出所: リクルートワークス研究所 Works Report: 未来予測2040 労働供給制約社会がやってくる

バックオフィス(BSM・HRM)領域のAI SaaS。AIにより業務そのものをなくし、生産性を改善

AI-powered

Approval Workflow

Accounting Automation

Compliance Automation

複雑・短期間・ミスできない

ルール、法律、業種・業態、社内事情に対応するための複雑な業務

毎月、短期間で大量の業務を完遂

それを「ミスなく」完遂

業務に高頻度で潜むアナログな手間

- 請求書を1枚1枚スキャンする手間
- 領収書をシステムに手入力する手間
- 仕訳を作成する手間
- カード明細と領収書との突合の手間
- 書類の情報を入力する手間

Bakuraku AI

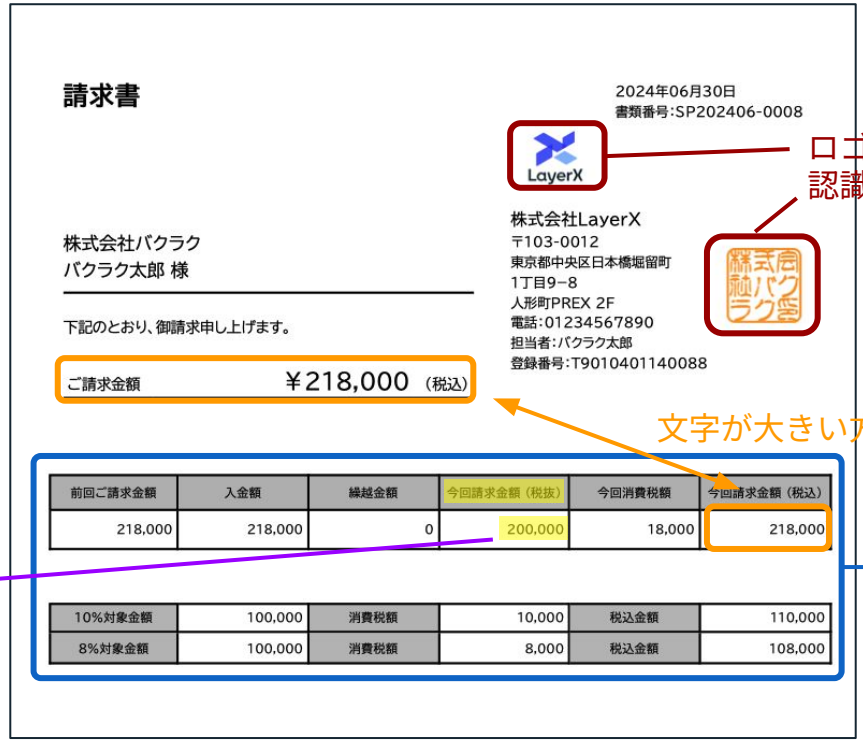
- **請求書をAIが自動分割して取り込み**
- **領収書のデータをAIが入力**
- **AIが過去に学習した仕訳を入力**
- **領収書とカード明細をAIが紐付け**
- **AIが書類種別を判定してラベル付け**

AIにより、アナログな手間を無くしていく

Introduction

ドキュメント解析技術とCV分野の繋がり

ドキュメント解析とComputer Vision分野との繋がり



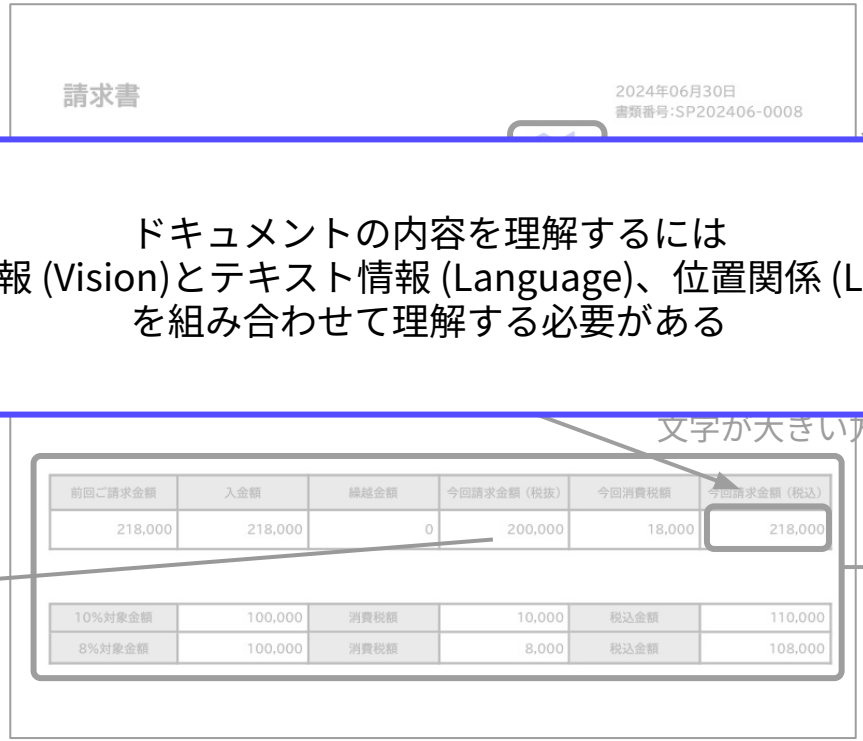
ロゴや印影・ハンコから
認識したい？

文字が大きい方を取りたい？

蛍光ペンを
引いてるところを
認識したい？

表形式で縦と横の
依存関係が混在

ドキュメント解析とComputer Vision分野との繋がり



ドキュメントの内容を理解するには
視覚情報 (Vision)とテキスト情報 (Language)、位置関係 (Layout)
を組み合わせて理解する必要がある

文字が大きい方を取りたい？

蛍光ペンを
引いてるところを
認識したい？

表形式で縦と横の
依存関係が混在

ドキュメント解析とComputer Vision分野との繋がり

PaddlePaddleでは、LLM以前からドキュメント解析を複合的な技術を組み合わせて行っている

PP-StructureV2: A Stronger Document Analysis System

Chenxia Li, Ruoyu Guo, Jun Zhou, Mengtao An,
Yuning Du, Lingfeng Zhu, Yi Liu, Xiaoguang Hu, Dianhai Yu
Baidu Inc.
{lichenxia, zhulingfeng}@baidu.com

Abstract

A large amount of document data exists in unstructured form such as raw images without any text information. Designing a practical document image analysis system is a meaningful but challenging task. In previous work, we proposed an intelligent document analysis system PP-Structure. In order to further upgrade the function and performance of PP-Structure, we propose PP-StructureV2 in this work, which contains two subsystems: Layout Information Extraction and Key Information Extraction. Firstly, we integrate Image Direction Correction module and Layout Restoration module to enhance the functionality of the system. Secondly, 8 practical strategies are utilized in PP-StructureV2 for better performance. For Layout Analysis model, we introduce ultra lightweight detector PP-PicoDet and knowledge distillation algorithm FGD for model lightweighting, which increased the inference speed by 11 times with comparable mAP. For Table Recognition model, we utilize PP-LCNet, CSP-PAN and SLAHead to optimize the backbone module, feature fusion module and decoding module, respectively, which improved the table structure accuracy by 6% with comparable inference speed. For Key Information Extraction model, we introduce VI-LayoutXLM which is a visual-feature independent LayoutXLM architecture, TB-YX sorting algorithm and U-DML knowledge distillation algorithm, which brought 2.8% and 9.1% improvement respectively on the Hmean of Semantic Entity Recognition and Relation Extraction tasks. All the above mentioned models and code are open-sourced in the GitHub repository PaddleOCR [1].

Document Layout Analysis can be regarded as an object detection task for document images in essence. The basic units such as titles, paragraphs, tables, and illustrations in the document are the objects needed to be detected and recognized. Layout-parser(Shen et al. 2021) is a unified toolkit for Deep Learning Based Document Image Analysis. VSR(Zhang et al. 2021) is proposed for layout analysis, which comes to state-of-the-art on PubLayNet dataset(Zhong, Tang, and Yepes 2019). In PP-Structure, we use PP-YOLOv2(Huang et al. 2021) to complete the layout analysis task, which is real-time on GPU devices. However, currently proposed models are not CPU-friendly and thus not conducive to deployment on CPUs or mobile devices.

Table Recognition is used to convert table images into editable Excel format files. The diversity of tables in document images, such as various rowspans and colspans and different text types, makes table recognition a hard task in document understanding. There are many table recognition methods, such as traditional algorithms based on heuristic rules and recently developed methods based on deep learning. Among them, the end-to-end method has received extensive attention due to the simplicity of the pipeline, which represent the table in HTML format and adopt Seq2Seq(Sutskever, Vinyals, and Le 2014) to predict the table structure, such as TableRec-RARE(Du et al. 2021b) in PP-Structure powered by PaddlePaddle(Ma et al. 2019). In TableMaster(Ye et al. 2021), transformer is used as the decoder, which achieves high accuracy, but brings huge computation cost.

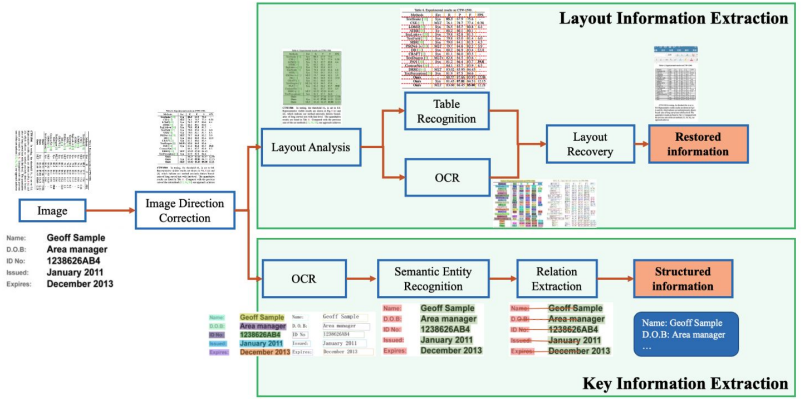


Figure 1: Framework of the proposed PP-StructureV2. It contains two subsystems: layout information extraction and key information extraction.

Chenxia Li, et al., *PP-structurev2: A stronger document analysis system*, arxiv, 2022.

ドキュメント解析とComputer Vision分野との繋がり

近年では、QwenなどのLVLMでDocument画像とBBoxも用いて学習されることもある。

Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution

Peng Wang* Shuai Bai* Sinan Tan* Shijie Wang* Zhihao Fan* Jinze Bai*†
Keqin Chen Xuejing Liu Jialin Wang Wenbin Ge Yang Fan Kai Dang Mengfei Du
Xuancheng Ren Rui Men Dayiheng Liu Chang Zhou Jingren Zhou Junyang Lin†
Qwen Team Alibaba Group

Abstract

We present the Qwen2-VL Series, an advanced upgrade of the previous Qwen-VL models that redefines the conventional predetermined-resolution approach in visual processing. Qwen2-VL introduces the Naive Dynamic Resolution mechanism, which enables the model to dynamically process images of varying resolutions into different numbers of visual tokens. This approach allows the model to generate more efficient and accurate visual representations, closely aligning with human perceptual processes. The model also integrates Multimodal Rotary Position Embedding (M-RoPE), facilitating the effective fusion of positional information across text, images, and videos. We employ a unified paradigm for processing both images and videos, enhancing the model's visual perception capabilities. To explore the potential of large multimodal models, Qwen2-VL investigates the scaling laws for large vision-language models (LVLMs). By scaling both the model size-with versions at 2B, 8B, and 72B parameters-and the amount of training data, the Qwen2-VL Series achieves highly competitive performance. Notably, the Qwen2-VL-72B model achieves results comparable to leading models such as GPT-4o and Claude3.5-Sonnet across various multimodal benchmarks, outperforming other generalist models. Code is available at https://github.com/QwenLM/Qwen2-VL.

1 Introduction

In the realm of artificial intelligence, Large Vision-Language Models (LVLMs) represent a significant leap forward, building upon the strong textual processing capabilities of traditional large language models. These advanced models now encompass the ability to interpret and analyze a broader spectrum of data, including images, audio, and video. This expansion of capabilities has transformed LVLMs into indispensable tools for tackling a variety of real-world challenges. Recognized for their unique capacity to condense extensive and intricate knowledge into functional representations, LVLMs are paving the way for more comprehensive cognitive systems. By integrating diverse data forms, LVLMs aim to more closely mimic the nuanced ways in which humans perceive and interact with their environment. This allows these models to provide a more accurate representation of how we engage with and perceive our environment

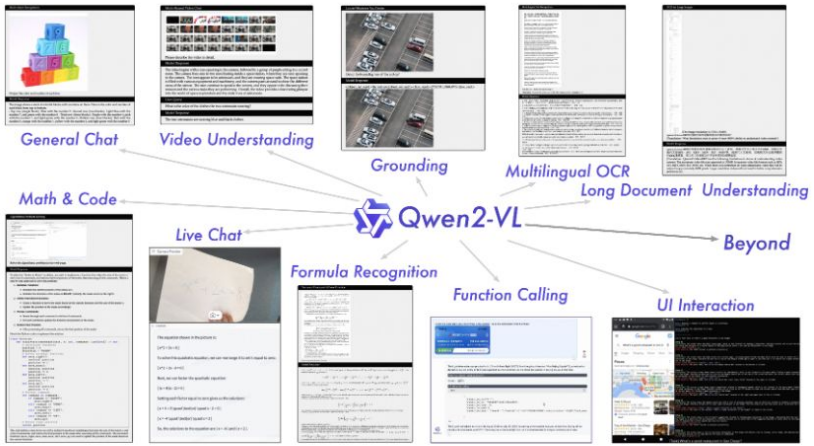


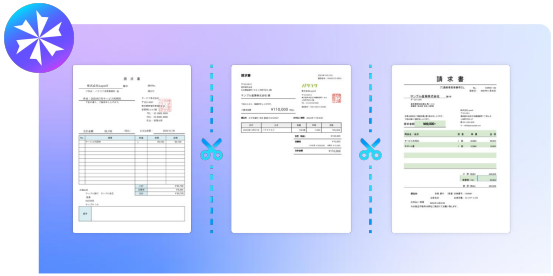
Figure 1: Qwen2-VL capabilities: Multilingual image text understanding, code/math reasoning, video analysis, live chat, agent potential, and more. See Appendix for details.

Peng Wnag, et al., *Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution*, arxiv, 2024

Bakuraku AI

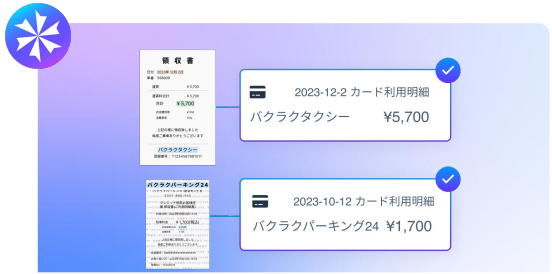
日常業務の中で自然にAIを活用いただけるような体験を提供

写真を撮影し、まとめて
アップロードするとAIがデータ入力



複数枚まとめてスキャンすると、
PDFファイルをAIが自動分割

請求書をアップロードした瞬間に、
AIが過去に学習した仕訳を入力



領収書をアップロードした瞬間にAIが
カード明細情報と照合適切な明細と紐付け

あらゆる種別の書類をまとめてアップロード
すると、AIが書類種別を判定し、ラベル付与

Bakuraku AIに関わる要素技術

- PDFや画像といった非構造化データからの項目抽出や分類
- 情報抽出されたデータや顧客の履歴データなどの構造化データを用いた推薦モデル

画像やPDFからの**項目抽出タスク**

各社の運用に合わせて
項目抽出した値を推薦

請求書に対して**仕訳を推薦**

画像やPDFの**書類分類**

領収書に対して**カード明細を推薦**

複数の請求書から抽出した情報から
尤もらしい**分割点を予測**

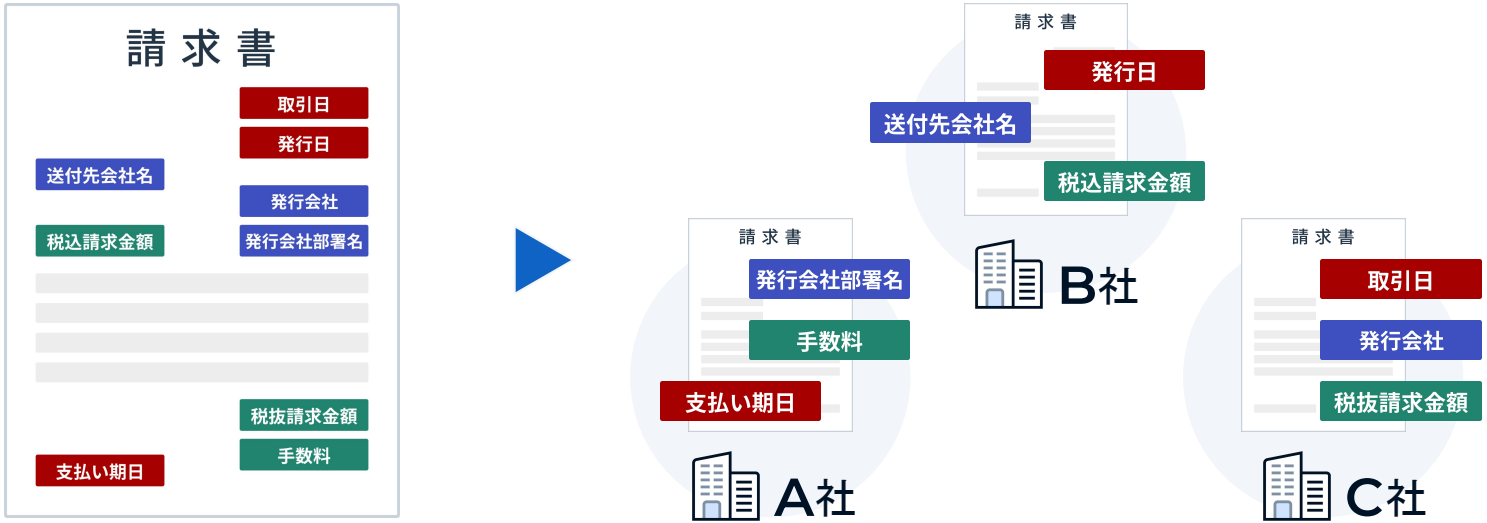
ドキュメント画像からの情報抽出

構造化データを用いた推薦モデル

パーソナライズドAI-OCR *特許出願準備中

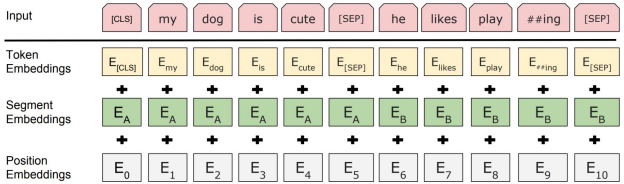
ユーザーに合わせて自動学習する次世代のAI-OCR
単純に情報抽出するだけではなく、項目抽出した値をパーソナライズすることを実現

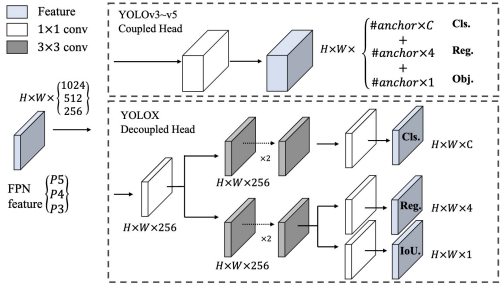
1 複数の値を同時に項目抽出 ▶ 2 お客様の選択を学習していき、運用に最適化していく

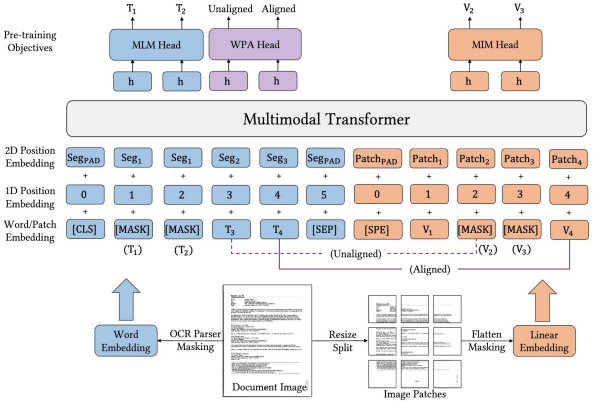


パーソナライズドAI-OCR (項目抽出モデル)

項目抽出部分はBERT系 (NLPモデル)や、Object Detection (CVモデル)系、
LayoutLM (マルチモーダルモデル)系などを複数検証







Jacov Deblin, et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, ACL, 2019

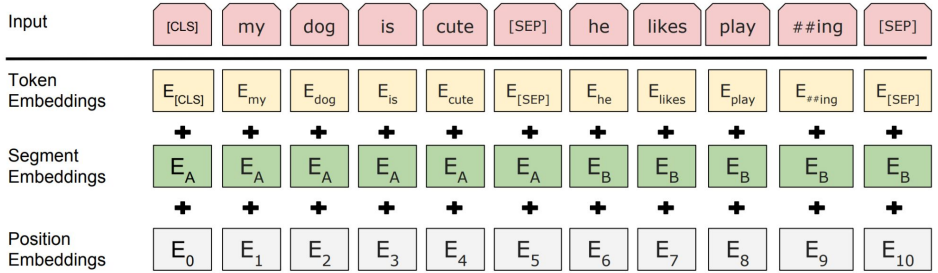
Zheng Ge, et al., *YOLOX: Exceeding YOLO Series in 2021*, arxiv, 2021

Yupan Huang, et al., *LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking*, ACM, 2022

パーソナライズドAI-OCR (項目抽出モデル)

RoBERTa

- モデル入力
  - OCRされた文書テキスト
- モデル出力
  - Tokenに対応するラベル (Token Classification)
- 選定理由
  - まずはシンプルな実装で ベースラインを作るのが重要

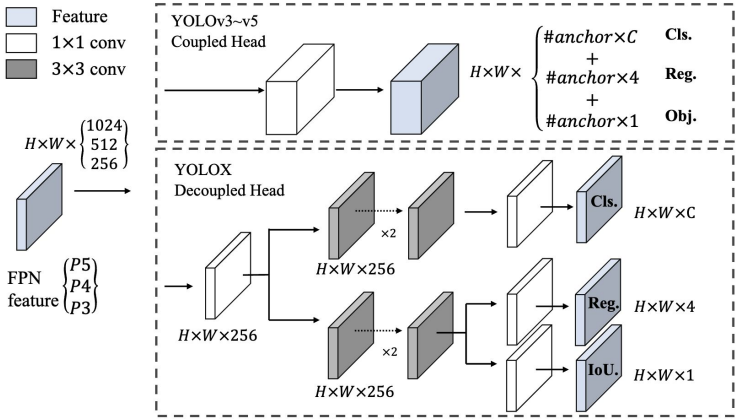


	20	21	22	23	24	25	26	27	28	29	30
token	。	合計	金額	¥	49	,	895	請求	書	様	御
label	others	others	others	others	paymentAmount	paymentAmount	paymentAmount	others	others	others	others

パーソナライズドAI-OCR (項目抽出モデル)

YOLOXなどのObject Detectionモデル

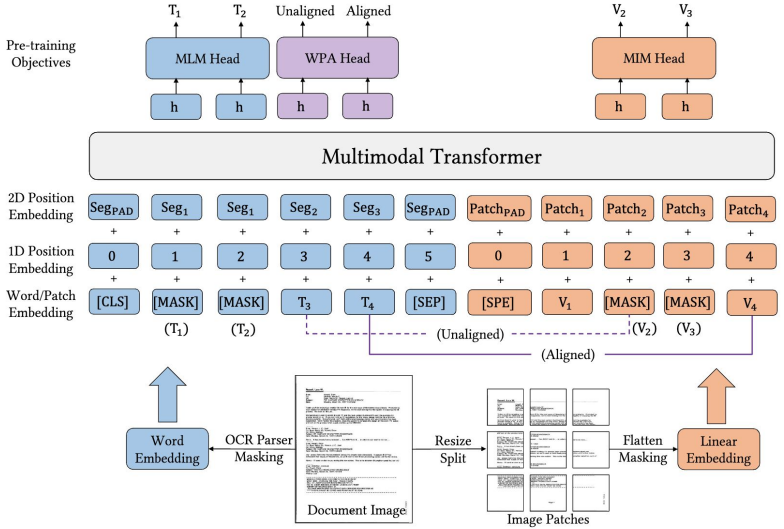
- モデル入力
  - PDFを画像化したもの
  - 携帯で撮った写真
- モデル出力
  - 欲しいラベルのBBox
  - OCR結果と突合して出力する
- 選定理由
  - 画像情報がどれほど重要かを検証



パーソナライズドAI-OCR (項目抽出モデル)

LayoutLMv3

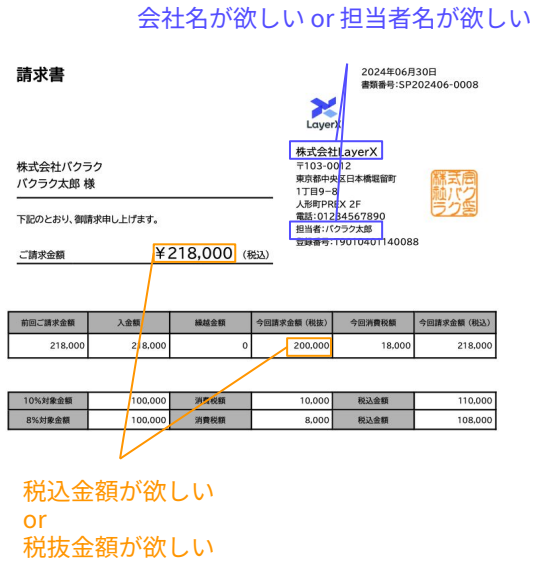
- モデル入力
  - OCRされた文書テキスト
  - PDFを画像化したもの
  - 携帯で撮った写真
- モデル出力
  - Tokenに対応するラベル (Token Classification)
- 選定理由
  - 当時のSoTAかつToken Classificationの枠組みで解ける



	20	21	22	23	24	25	26	27	28	29	30
token	。	合計	金額	¥	49	,	895	請求	書	様	御
label	others	others	others	others	paymentAmount	paymentAmount	paymentAmount	others	others	others	others

パーソナライズドAI-OCR (推薦モデル)

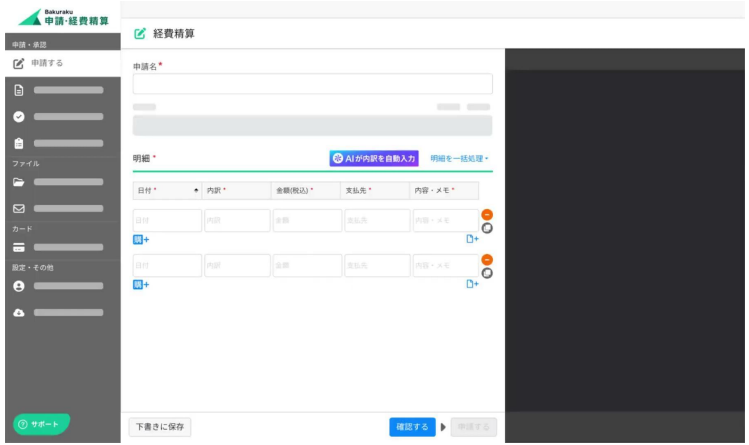
最終的にお客様が欲しいのは、「誰に対して」「いつ」「いくら支払った」といった情報。
同じ書類であっても、お客様の運用によって変わることがあるため、過去の入力履歴を活用。



(余談) Why not LLM ?

LayerXでは、AIを用いた際の体験「**AI-UX**」を重視しており、
間違えた際の**気づきやすさ**、**修正しやすさ**、**抽出速度**などからシンプルなモデルを使用。
(どこからその値を抽出したか、などはまだ生成モデルより既存モデルの方が強い。)
自社ドメインの大量データで学習する際は、精度的にも既存モデルの方が高いことも多い。

また、パーソナライズ部分などもICLより既存の推薦技術の方がシンプルで扱いやすい。



証憑マッチング・仕訳推薦

AI-OCRで抽出した取引先名や金額・日付などの情報を元に

- 「クレジットカードの明細」と「領収書ファイル」の紐付けや
- 「仕訳」と「請求書」の紐付けを自動化。

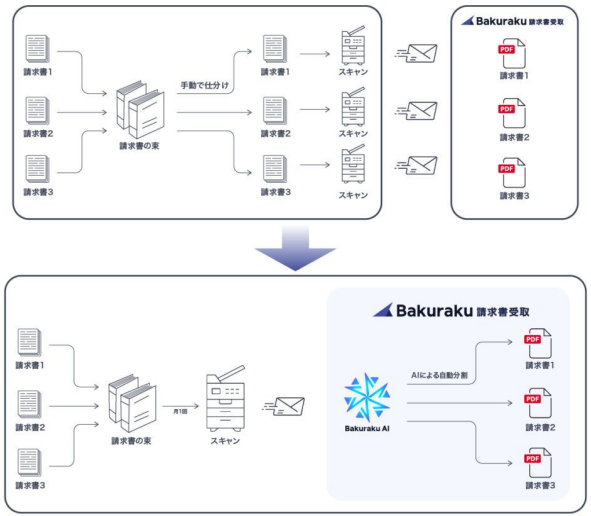
未処理のカード明細を候補として推薦

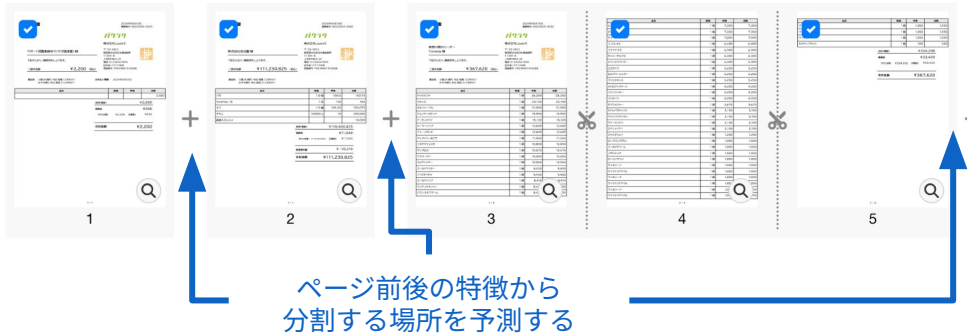
過去の仕訳情報を候補として推薦

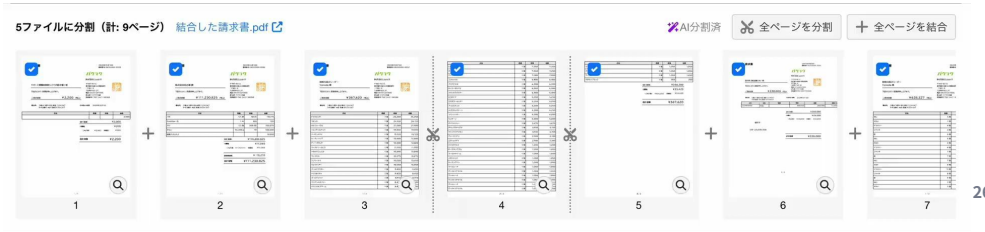
PDF分割

請求書を取引ごとに管理するため、取引ごとにお客様が書類を分割してスキャンすることが必要だった

AI-OCRで抽出した情報を特徴に、分割点を予測。







実データにおける課題

パーソナライズドAI-OCRを例に

訓練ラベルの設計

実際のプロダクトでのユースケースを加味する必要があるため、**ドメインにDeepDiveする必要**がある

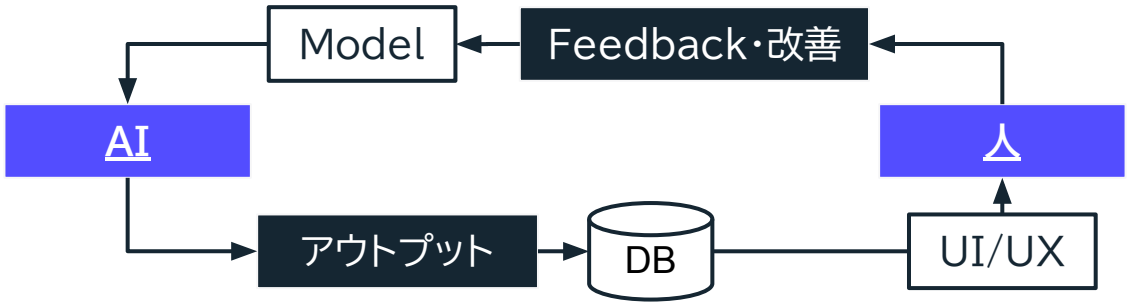
例) パーソナライズドAI-OCRの場合、以下のようなことを加味したラベル設計をしなければいけない。

- 領収書の場合、法律的に屋号でも運営会社名でも取引先として抽出して大丈夫。
- 新幹線の領収書は、購入日なのか乗車日を取りたいかがサービスによって変わる。

フィードバックサイクルの設計が重要

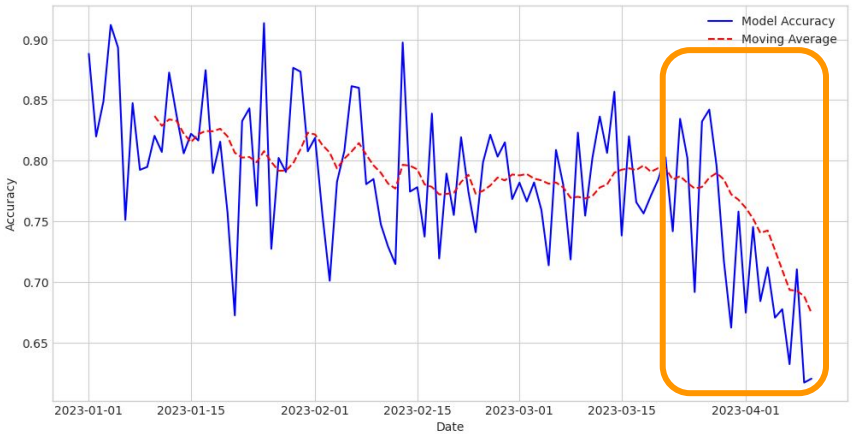
- お客様が増えていく中で、書類の多様性なども日々増えていくが、
  全ての書類に対してアノテーションできるわけではない。
  - 間違えやすい書類をどのように集めてアノテーションに回すか。
  - お客様のFBをどのように次の学習に活用するか。
- 修正してくれるお客様とそうでないお客様が存在する中、ラベルノイズの影響をどう減らすか。

これらをプロダクトをリリースする前に設計し、必要なログなどを仕込む必要がある。
機械学習モデルの開発は**デプロイして終わりではない**。

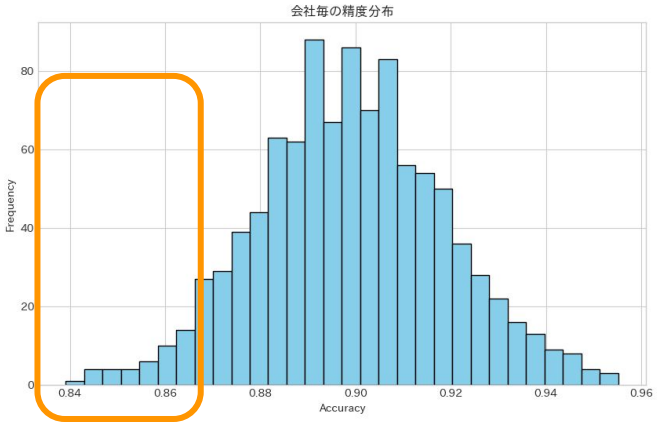


デプロイした後も常にモニタリングし、再設計し続ける

- ラベルの作り方に間違った仮定が入ってなかったか
- 想定していなかった運用のお客様が増えていないか
- 実際にお客様に話を聞きにいくのも重要



急激な精度変化の検知



精度が比較的低いお客様で
何が起きているのかの分析

※グラフはダミーデータです

実際に出てきているユースケース

- 経費精算の際に台紙に領収書を複数貼って提出する運用。
  - AI-OCRは一つの画像に複数の取引先名が入っていることを想定していないため、精度劣化。
- 一つの請求書の合計金額ではなく、明細の金額を用いて複数の仕訳に分割したい。
  - AI-OCRは合計金額だけを読み取っていたが、明細の金額も読み取る必要が出てきた。

請求書

株式会社御中	クルクル商店

請求明細表

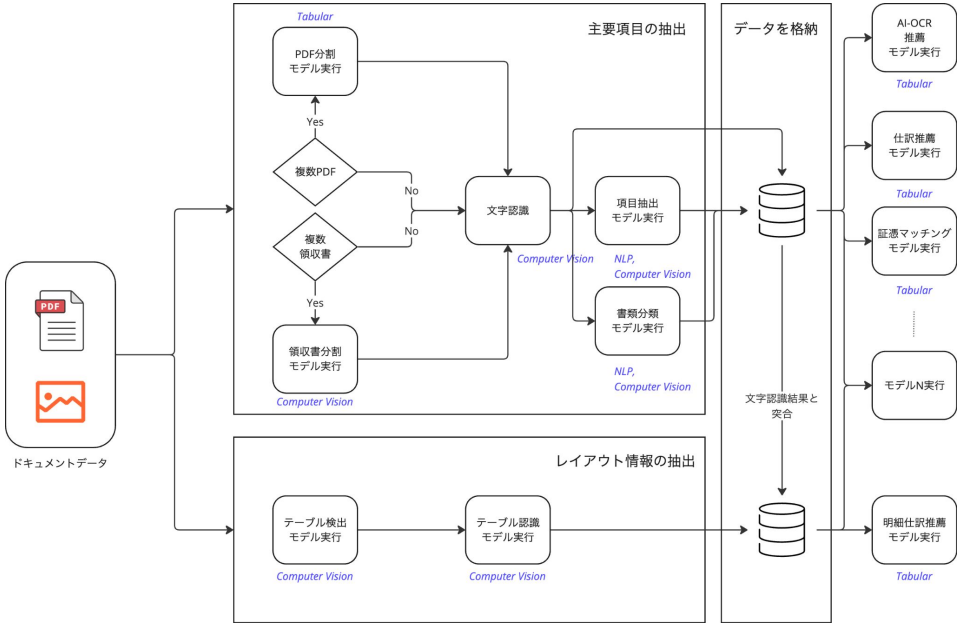
ご利用年月日	問い合わせ番号	品名	請求金額	税率(%)
2024/04/30	73150452	トイレットペーパー　新宿店	481,961 円	10
2024/05/31	73150454	トイレットペーパー　名古屋店	139,306 円	10
2024/06/30	73150456	トイレットペーパー　札幌店	243,745 円	10
2024/07/31	73150458	トイレットペーパー　福岡店	73,780 円	10
2024/08/31	73150460	備品　大阪店様	210,215 円	8
2024/09/30	73160890	備品　上野店様	150,437 円	8
2024/10/31	73190221	備品　石川店様	128,366 円	8
2024/11/30	73109442	備品　広島店様	300,773 円	8
			小計	1,728,583
			消費税等(10%)	86,501
			消費税等(8%)	69,086
				1,884,170

コンゴトモヨロシクオネガイシマス

今後の展望

ドキュメント解析パイプラインの改善

実際に出てきているユースケースを元にデータ抽出パイプラインを再設計し、
後段で様々なユースケースでの活用を行っていきたい。



まとめ

LayerXでは様々な技術を組み合わせてお客様の課題を解決しています

- ドキュメントという非構造化データを起点に
  CVやNLP、推薦などの複合的な技術を用いて、お客様の課題を解決しています。
- モデルを作成する技術はもちろんですが、継続的に改善をするために、
  ログの設計やFBサイクルの設計などのMLOps的要素をモデル作成時から考える必要がある。
- 今後、さらに多様なユースケースを解決していくために、
  様々な技術を用いてお客様の体験を「バクラク」にしていきます。

[宣伝] LayerX Machine Learning勉強会

LayerXでは毎週機械学習関連の勉強会を行っており、内容を全て公開しています。
話題になった論文などの紹介を行っているので、ぜひご覧ください！

メインTOPIC

[論文] The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, David Ha

概要

Sakana AIが提案する論文作成のような科学的発見プロセスを完全に自動化する包括的なフレームワーク

- 大規模言語モデル(LLM)を活用し、研究のアイデア生成から論文執筆・査読までを自動化
- 1本の論文生成にかかるコストはわずか$15程度と極めて効率的
- 人間の査読者と同等レベルの自動査読システムを実現
- 機械学習の3つの異なる分野（拡散モデル、言語モデル、学習ダイナミクス）での有効性を実証

Introduction