視覚的質問応答による文書情報抽出における同時多項目推論

Mengsay Loem　保坂 大樹
Sansan 株式会社
{mengsay.loem, hosaka}@sansan.com

概要

視覚的質問応答は、文書画像から構造化された情報を抽出する上で有望な手法である。しかし、従来の手法では、抽出対象となる各項目を個別に質問応答する方式で実現されることが多く、項目間に内在する依存関係を十分に活用できていない可能性がある。本研究では、関連性のある複数項目を同時に推論する手法を検討し、その有効性を実証する。実世界のデータセットを用いて評価した結果、相互依存性が高い項目群において、同時に推論する手法は従来の手法を大幅に上回る抽出精度を示し、相互依存性が低い項目に対しても同等の水準を維持できることを示した。また、項目数や項目間の関連の強さが抽出精度に及ぼす影響を分析し、視覚的質問応答に基づく情報抽出を効果的に実現するための知見を提示する。

1　序論

視覚言語モデル (Vision Language Models, VLMs) [1, 2, 3] の進歩に伴い、視覚的質問応答 (Visual Question Answering, VQA) [4] は、文書画像から構造化情報を抽出する手法として注目されている [5, 6, 7]。特に、プロンプトベースの VQA (PromptVQA) では、ユーザが自然言語による問いかけを通じて文脈に応じた回答を得ることが可能である。そのため、複雑なレイアウトと高密度なテキストが混在する領収書や請求書などのビジュアルリッチな文書に対しても、多様な形式の質問を追加学習を要せずに行える点が大きな強みとなっている [8, 9]。こうした利点により、VQA は情報抽出 (Information Extraction, IE) における拡張性の高いアプローチとして注目を集めている。

しかし、既存の VQA ベース IE 手法は、図 1(左) に示す、独立単項目 VQA (Independent Single-Item VQA, InSiVQA) 方式を採用することが多い [10, 11]。すなわち、必要な情報ごとに個別の質問を投げる方法である。これは直接的かつ単純な手法である一方、各

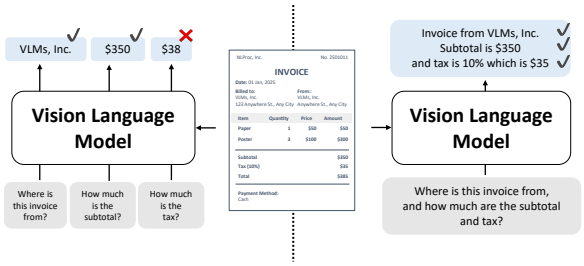

図 1: 独立単項目 VQA (InSiVQA, 左) 方式と複数項目同時推論 VQA (JoMiVQA, 右) 方式の概要図。

項目間の関係を無視してしまう欠点がある。例えば、請求書の合計金額をより正確に確認するためには、関連する行項目や課税対象額と税額など、他の関連情報を同時に考慮することが有用な可能性がある。しかし、InSiVQA ではこうした依存関係を活用できず、その結果、潜在的な性能向上が十分に引き出せない懸念がある。

この問題に対し、近年の研究では複数の質問の同時推論が模索されている。例えば、Tang ら [12] は、単一推論ステップで複数質問に対応可能な新たなモデルアーキテクチャと、それに適した事前学習および学習可能なプロンプト戦略を組み合わせることで、複数質問を同時に考慮する手法を提案している。これにより性能改善が報告されているものの、その成果がモデルアーキテクチャや追加の事前学習、あるいは相互依存する項目を同時に推論する本質的な効果のどれに起因するのかは明確でない。

そこで本研究では、モデルのアーキテクチャ改修や追加事前学習を行わずに、依存関係を有する複数の項目を同時に推論することが、VQA ベース IE の性能向上に役割を果たし得るかを検証する。具体的には、PromptVQA において複数の質問を同時に VLM へ入力する Joint Multi-Item VQA (JoMiVQA) を提案する (図 1、右)。本手法を同一の VLM 上で InSiVQA と比較することで、相互依存する項目を一括して推論することによる効果と、モデルアーキテクチャや事前学習による影響を切り離し、依存関係

を考慮することがいつ、どのように有用であるかを明らかにする。

実験では、実世界のレシートデータセットを用い、0-shot および few-shot 設定の下、複数の VLM を用いて評価を行った。その結果、依存関係を持つ項目（例えば、税額と小計）において、JoMiVQA は一貫して InSiVQA を上回る性能を示すことが確認された。一方、依存性の低い項目や識別が容易な項目については、依存性の高い項目と比較して両手法間の性能差が縮小することも確認された。さらに、回帰分析を導入することで、項目間の相互依存性を定量化し、性能への影響を実証的に明らかにした。また、項目数や依存関係の要因が精度に与える影響を考察することで、VQA ベース IE の最適化に向けた実用的な知見を提供する。

2 関連研究

近年の研究では、文書画像中のテキスト、レイアウト、視覚的特徴を統合的に扱う枠組みが進展し、情報抽出、表構造解析などの多様なタスクで大きな性能向上が報告されている。例えば、LayoutLM 系モデル [13, 14, 15] や LaTr [10] は、文書上のテキスト情報と空間的特徴を組み合わせることで文書構造を効果的に捉え、高精度な情報抽出を実現している。

一方、画像と言語の統合処理をする VLM を用いた PromptVQA が登場し、大規模な画像とテキストデータにより事前学習された CLIP [16]、Qwen VL [17, 2]、Llama3 [3] などでは、0-shot や few-shot の設定で堅牢な性能が示されている。PromptVQA は、例えば「この領収書の合計金額は何か？」や「この請求書の明細を一覧せよ」といった自然言語での指示に直接応答できる。そのため、レイアウト特徴を強調したモデルなどの従来手法とは異なり、追加学習の手間を最小限に抑えつつユーザニーズに合わせた多様な問い合わせを行える利点を持つ [8, 9]。

しかし、従来手法は InSiVQA 方式を踏襲することが多く、個別項目を独立した質問で処理するため、項目間依存性を十分に引き出せない問題がある。一部の研究 [12] は複数項目を同時に扱う手法に言及しているものの、同時推論が独立推論に対してどの程度本質的な利点をもたらすかについては未解明の点が多い。本研究では、この未解明領域に踏み込み、アーキテクチャや事前学習上の工夫とは切り離した形で、複数項目を同時に考慮することによる有効性を検証する。

3 InSiVQA と JoMiVQA の比較

本節では、InSiVQA と JoMiVQA を比較するための実験設定および結果を示す。複数の項目を同時に推論することで、各項目を独立に推論する場合より精度が向上するかを検証する。

3.1 VQA 用プロンプト戦略と評価

0-shot と 4-shot の 2 種類の推論条件で InSiVQA および JoMiVQA を評価する。モデルへの指示文には JSON 形式での出力指定を含め、正規表現を用いて評価対象項目の出力を抽出する[1)]。本研究では、抽出された文字列と正解アノテーションが完全一致した場合を正解と判定する。ただし、金額関連の項目については通貨記号やカンマを取り除いた上で比較を行い、情報が存在しない場合はモデルの出力が空文字列であれば正解とみなす。

InSiVQA　この方式では、合計金額や会社名などの抽出対象情報をすべて独立した質問として与える。具体的には、以下のような質問を用いる。

> What is the {`target information`} on this receipt?

JoMiVQA　この方式では、相互に関連する複数の項目を同時に問い合わせる。モデルが複数の関連項目を同時に考慮し、文脈的・意味的関連性を活用することで、精度の向上が期待される。例えば、以下のような質問を用いる。

> What are the `company name`, `itemized prices`, and the `total amount` on this receipt?

3.2 使用モデルおよびデータセット

評価には、Llama-3.2-11B-Vision-Instruct[2)] および Qwen2-VL-2B-Instruct[3)] の 2 種類の VLM を用いる。また、より実世界に近いシナリオで性能評価を行うため、以下の 2 つのベンチマークを利用する。

CORDv2　データセットは 800 件の訓練データと、開発・テスト用に各 100 件のデータから構成される [18]。本データセットには、実世界のスキャンレシートが含まれ、項目名・数量・小計・税額など、多階層のセマンティックラベルが付与されている。本実験では主に金額に関する項目を抽出対象とし、Subtotal (小計) 、Tax (税額) 、Service (サービス料) 、Total (合計金額) 、CreditCard (クレジットカード支払

1) 抽出が失敗する場合は著者が目視で評価する。
2) https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
3) https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct

表 1: CORDv2 データセットにおける JoMiVQA と InSiVQA の正解率比較。

項目名	0-shot						4-shot					
	Llama-3.2-11B-Vision-Instruct			Qwen2-VL-2B-Instruct			Llama-3.2-11B-Vision-Instruct			Qwen2-VL-2B-Instruct		
	InSiVQA	JoMiVQA	差	InSiVQA	JoMiVQA	差	InSiVQA	JoMiVQA	差	InSiVQA	JoMiVQA	差
Subtotal	0.69	0.72	+0.03	0.92	0.93	+0.01	0.68	0.67	-0.01	0.90	0.95	**+0.05**
Total	0.74	0.75	+0.01	0.96	0.98	+0.02	0.69	0.69	0.00	0.94	0.98	+0.04
Tax	0.65	0.79	**+0.14**	0.79	0.89	**+0.11**	0.16	0.29	**+0.13**	0.84	0.97	**+0.13**
CreditCard	0.93	0.93	0.00	0.87	0.88	+0.01	0.73	0.80	**+0.07**	0.80	0.80	0.00
Quantity	0.60	0.81	**+0.21**	0.76	0.96	**+0.20**	0.52	0.52	0.00	1.00	1.00	0.00
Cash	0.67	0.93	**+0.26**	0.87	0.92	**+0.05**	0.82	0.83	+0.01	0.78	0.97	**+0.19**
Change	0.89	0.94	**+0.05**	0.92	0.98	**+0.06**	0.71	0.77	**+0.06**	0.92	0.96	+0.04
Service	0.83	0.92	**+0.09**	0.67	0.75	**+0.08**	0.25	0.25	0.00	0.92	0.92	0.00
全項目	0.40	0.71	**+0.31**	0.77	0.83	**+0.06**	0.45	0.64	**+0.19**	0.69	0.89	**+0.20**

い額) 、Cash (現金支払い額) 、Change (釣銭) などを対象とする。評価には、テストセットを用いる。

SROIE Task 3 は SROIE 2019 コンペティションで提示されたタスクであり、スキャンレシートから会社名、住所、取引日、合計金額などの情報を抽出することが求められる [19]。評価には 347 件のデータを含むテストセットを用いる。

表 2: SROIE データセットにおける正解率比較。

項目名	Llama-3.2-11B-Vision-Instruct			Qwen2-VL-2B-Instruct		
	InSiVQA	JoMiVQA	差	InSiVQA	JoMiVQA	差
Company name	0.92	0.92	0.00	1.00	1.00	0.00
Address	0.98	0.95	-0.03	1.00	1.00	0.00
Date	1.00	1.00	0.00	0.99	0.99	0.00
Total Amount	0.90	0.91	+0.01	0.95	0.96	+0.01
全項目	0.81	0.83	+0.02	0.94	0.95	+0.01

3.3 結果

表 1 は、CORDv2 のテストセットにおける InSiVQA と JoMiVQA[4] の正解率を比較した結果を示している。0-shot の設定では、全項目の正解率で、JoMiVQA は InSiVQA に対して顕著な改善を示している。特に、Tax (税額) や Cash (現金支払額) 、Quantity (数量) など、金額や数量を扱う相互依存性の高い項目では、JoMiVQA による大幅な精度向上が確認された。一方で、CreditCard のように独立性が高く単体でも容易に推定可能な項目では、両方式間の差は小さく、項目間依存関係が低い場合には JoMiVQA の恩恵が限定的であることが示唆される。また、4-shot[5] では、少数ながら事例を提示することで、両方式間の差は全体的に縮小する傾向がある。しかし、依然として Tax や Cash などの依存関係が強い項目では、

4) JoMiVQA における同時に推論する項目は、各テストケースの正解アノテーションが付いている全項目とする。

5) 事例は訓練データからランダムに選択される。

表 3: JoMiVQA の出力例。

対象項目	モデル出力の一部
Tax	The receipt shows a total amount of 51,300 and a tax price of 10%. We can calculate the tax price as follows: Tax Price = Total Amount x Tax Rate = 51,300 x 0.10 = 5,130
Change	The receipt shows a total amount of 80,500 and a cash payment of 100,000. To calculate the change, we subtract the total from the payment amount: 100,000 - 80,500 = 19,500. This calculation reveals that the change amount is 19,500.

JoMiVQA が InSiVQA を大きく上回る正解率を維持している。表 2 は SROIE Task 3 データセットにおける結果を示している。CORDv2 に対し、JoMiVQA と InSiVQA の間に有意な精度差はほとんど見られなかった。SROIE では、会社名や住所、合計金額といった項目が相対的に独立性が高く、単独で容易に識別可能であることが、複数項目同時推論の恩恵を限定的なものにしていると考えられる。表 3 は、JoMiVQA による推論結果の例を示している。

これらの結果は、複数項目を一括して推論する方式が、特に項目間に依存性が存在する場合に大きな性能改善をもたらすことを示している。

4 依存性が JoMiVQA に与える影響

本節では、項目間の相互依存度を定量化し、複数項目を同時に推論する JoMiVQA が性能向上に寄与する条件を検証する[6]。

4.1 項目間依存関係の定義

分析対象として、CORDv2 データセットの 3 項目組 (x, y, z) を考える。ここで x は検証対象の項目 (例: Tax) であり、y, z は x と関連がある可能性のある項目 (例: Subtotal, Total) である。この 3 項目組に

6) モデルに入力する少数事例の偏りによる影響を排除するために、以降の実験においては 0-shot で実施される。

表 4: 対象項目と関連項目間の依存関係による、JoMiVQA と InSiVQA 間の正解率向上効果の比較。

対象項目	同時推論項目	R^2	サンプル数	Llama-3.2-11B-Vision-Instruct			Qwen2-VL-2B-Instruct		
				InSiVQA	JoMiVQA	差	InSiVQA	JoMiVQA	差
Tax	Subtotal, Total	0.99	80	0.60	0.80	**+0.20**	0.69	0.80	**+0.11**
	Change, Menu Types	0.05		0.60	0.59	-0.01	0.69	0.68	-0.01
Cash	Total, Change	0.95	100	0.72	0.81	**+0.09**	0.71	0.86	**+0.15**
	Change, Menu Types	0.02		0.72	0.69	-0.03	0.71	0.60	-0.11
Change	Total, Cash	0.98	100	0.84	0.90	**+0.06**	0.83	0.92	**+0.09**
	Subtotal, Tax	0.04		0.84	0.82	-0.02	0.83	0.84	+0.01

対し、訓練セットを用いて以下の重回帰モデルを適用する。

$$x = c_1 y + c_2 z + b$$

ここで、c_1, c_2, b は回帰係数および切片である。決定係数 (R^2) によって、x が y と z によりどの程度説明可能かを測定する。R^2 が高いほど、x が y, z に強く依存していることを意味する。本実験では、$R^2 \geq 0.9$ の場合を High-R^2 と定義し、x が y, z を用いて高精度に推定可能な強い依存関係があることを示す。$R^2 \leq 0.1$ の場合を Low-R^2 と定義し、x が y, z とほとんど無関係であり、依存性がないか極めて低いことを示す。仮説として、High-R^2 の場合は、JoMiVQA が InSiVQA に比べて有意な精度向上を示すと考えられ、Low-R^2 では項目間依存がほぼ無いため、JoMiVQA による有意な改善は期待できない。

4.2 結果

表 4 は、上記の分析結果を示す。Llama-3.2-11B-Vision-Instruct および Qwen2-VL-2B-Instruct のモデルについて、High-R^2 では JoMiVQA が InSiVQA を大きく上回る性能向上を示し、強い依存関係が複数項目同時推論による精度改善に直接寄与していることが確認された。一方、Low-R^2 では、両手法間にほとんど性能差が見られず、依存性が低い項目では同時推論の利点が限定的であることが明らかとなった。

これらの知見は、前節までの観察を補強している。すなわち、複数項目を同時に考慮する推論手法の有効性は、項目間の依存関係によって決定的に左右される。相互依存性の定量化により、JoMiVQA が特に有効となる条件を明示でき、実用的な判断基準として活用可能である。

5 質問項目数が与える影響

実世界の応用では、複数の情報項目を同時に抽出する必要がある。JoMiVQA の拡張性と実用性を評価するため、1 つの質問内で対象とする項目数を 2 から 6 まで変化させた実験を行う[7]。評価では、指定したすべての対象項目が正しく抽出できた場合のみ正解とみなす。

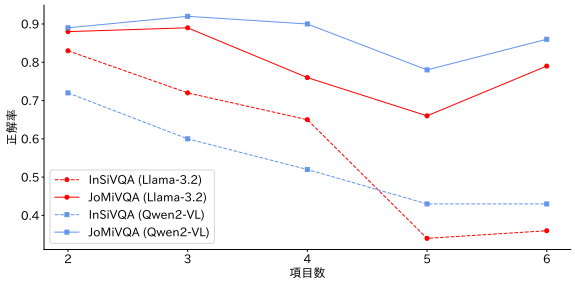

図 2: 項目数と正解率の変化。

図 2 は、InSiVQA と JoMiVQA の正解率の推移を示す。全ての項目数条件下で、JoMiVQA は両モデルにおいて InSiVQA を上回る精度を維持した。InSiVQA は要求項目数の増加とともに精度が急激に低下しており、項目数が増えるにつれて誤り発生の確率が累積的に高まることが一因と考えられる。一方、JoMiVQA は項目数が増加しても相対的に高い精度を保ち、タスクの複雑化に対しても堅牢な性能を示した。

6 結論

本研究は、VQA に基づく情報抽出において、相互依存する複数の項目を同時に推論する JoMiVQA が性能と頑健性に与える影響を考察した。相互依存度が高い項目群に対して、JoMiVQA は InSiVQA を大きく上回る精度向上を示すことを明らかにした。また、抽出項目数が増しても、JoMiVQA は堅牢な性能を維持し、単純な独立抽出手法に比べてスケーラビリティと効率性で優位性を発揮することが分かった。さらに、本研究では項目間の依存関係を定量化し、相互依存性が高い場合にこそ同時推論の効果が顕著になることを示した。

7) CORDv2 の開発とテストセットから正解アノテーションが付いている項目数でグループ化し、評価セットとする。

参考文献

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PMLR, 2021.

[2] 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, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. **arXiv preprint arXiv:2409.12191**, 2024.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and Amy Yang et al. The llama 3 herd of models, 2024.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In **Proceedings of the IEEE international conference on computer vision**, pp. 2425–2433, 2015.

[5] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pp. 2200–2209, January 2021.

[6] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, No. 15, pp. 13878–13888, May 2021.

[7] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. Docformerv2: Local features for document understanding. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, No. 2, pp. 709–718, Mar. 2024.

[8] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In **2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 14974–14983, 2023.

[9] Yihao Ding, Zhe Huang, Runlin Wang, YanHang Zhang, Xianru Chen, Yuzhong Ma, Hyunsuk Chung, and Soyeon Caren Han. V-Doc : Visual questions answers with Documents . In **2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 21460–21466, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.

[10] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. Latr: Layout-aware transformer for scene-text vqa. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 16548–16558, 2022.

[11] Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In **Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16**, pp. 715–732. Springer, 2020.

[12] Peng Tang, Srikar Appalaraju, R. Manmatha, Yusheng Xie, and Vijay Mahadevan. Multiple-question multiple-answer text-VQA. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)**, pp. 73–88, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[13] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining**, KDD ' 20, p. 1192–1200. ACM, August 2020.

[14] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. **arXiv preprint arXiv:2012.14740**, 2020.

[15] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In **Proceedings of the 30th ACM International Conference on Multimedia**, pp. 4083–4091, 2022.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[17] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[18] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. 2019.

[19] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In **2019 International Conference on Document Analysis and Recognition (ICDAR)**. IEEE, September 2019.