表画像入力による表構造解析手法の調査

A survey on table structure recognition methods using table image inputs

納田朋享
Tomoaki Noda
岡山大学太田研究室
Ohta Laboratory, Okayama University

概要 様々な文書で数値をまとめるために表が用いられており，表からより視覚的に優れたグラフに自動変換する研究が行われている．このような研究のためには，まず様々な形式で書かれる表の構造を解析する必要がある．また，表画像を入力とした表構造解析ができればその適用対象は広い．そこで本稿では，表画像入力による表構造解析手法を調査する．

1 はじめに

表は様々な文書で用いられる．例えば，学術論文では実験結果を表にまとめることが多いが，実験結果を一目で理解するには表より視覚的に優れたグラフが適している．そのため，文書中の表をグラフへ自動変換する研究が行われている[1]．このような表を活用した研究では，まず様々な形式がある表の構造を解析する必要がある．また表画像入力に対応できればより多くの文書を扱えるようになる．そこで本稿では，表画像入力による表構造解析手法について ICDAR 2021 Competition on Scientific Table Image Recognition to LaTeX[2] においてベースラインとして用いられた2つの手法について調査する．

2 表画像入力による表構造解析

2.1 問題定義

文献[2]における表画像入力による表構造解析とは，表画像が入力されたときに罫線や配置などの構造情報を含むトークンの系列をLaTeX形式で生成することである．なお，数値などセルの中の文字列は「CELL」というトークンで表される．

表画像入力による表構造解析の例を図1に示す．赤色で囲まれた表画像が入力されたときに水色で囲まれたLaTeX形式のトークンの系列を出力する．なお，表画像とトークンの系列の対応のため色付けしている．

2.2 Dengらの手法

Dengら[3]は，Convolutional Neural Network（CNN）と Long Short Term Memory（LSTM）[4]を用いて画像を入力としてマークアップ形式のトークンの系列を生成する手法を提案した．Dengらの手法の概要を図2に示す．

まず，画像を入力として特徴を抽出する．特徴抽出にはCNNを用いており，画像を入力としてチャネル数D，

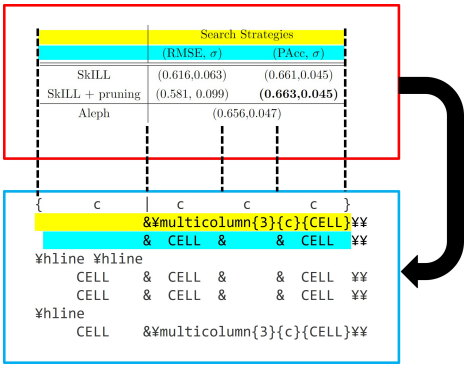

図1: 表構造解析の例[*1]

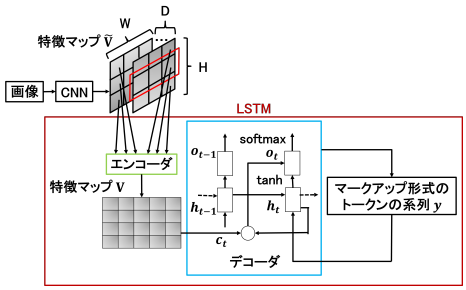

図2: Dengらの画像からマークアップ形式のトークンの系列を生成する手法の概要[3]

特徴マップの高さ（行）Hと幅（列）Wからなるサイズ$D \times H \times W$の特徴マップ$\tilde{\mathbf{V}}$ を生成することで画像の特徴を抽出する．

その後，抽出した特徴を用いてマークアップ形式のトークンの系列を生成する．マークアップ形式のトークンの系列の生成には Recurrent Neural Network（RNN）の一種であり長期的な依存関係を学習できるLSTMを用いている． LSTMはエンコーダとデコーダからなる．

エンコーダでは，抽出した特徴マップ$\tilde{\mathbf{V}}$を行ごとに入力することで新しい特徴マップ$\mathbf{V}$を生成する．

デコーダでは，特徴マップ$\mathbf{V}$に基づいてマークアップ形式のトークンの系列yを生成する．このデコーダは，履歴と注釈を用いて式（1）に示す次のトークンを与える条件付き言語モデルとして学習する．

$$p(y_{t+1}|y_1, ..., y_t, \mathbf{V}) = \text{softmax}(\mathbf{W}^{out}\mathbf{o}_t) \quad (1)$$

ここで，y_tは時刻tに出力されたマークアップ形式のトークン，$\mathbf{o}_t = \tanh(\mathbf{W}^c[\mathbf{h}_t; \mathbf{c}_t])$で定義

[*1] https://competitions.codalab.org/competitions/26979

された値であり， $\mathbf{W}^{out}$, $\mathbf{W}^{c}$は学習によって作られた重み付けのためのパラメータである．また $\mathbf{h}_t = \mathrm{LSTM}(\mathbf{h}_{t-1}, [y_{t-1}; \mathbf{o}_{t-1}])$と定義され， $\mathbf{h}_t$はデコードの履歴をまとめるために使用される． $\mathbf{c}_t$は画像の特徴の期待値で表されるコンテキスト情報であり，マークアップ形式のトークンの系列を生成するときに画像の次の位置を追跡できるかに依存した値である．

2.3 Fengらの手法

Fengら[5]は， Residual Networks（ResNet）[6]とTransformer[7]を用いて画像を入力としてトークンの系列を生成する手法を提案した． Fengらの手法の概要を図3に示す．

まず，入力画像の特徴を抽出する．特徴抽出にはCNNの一種であり，ネットワークの層をショートカットできる特徴をもつ ResNetを用いている． 101層からなるResNet-101の最初の4つの層を特徴抽出モジュールとして用い， 96×96画素に変換された入力画像から $6 \times 6 \times 1,024$のサイズの特徴マップを作成することで画像の特徴を抽出する．

その後，抽出した特徴を用いて構造情報のトークンの系列を生成する．トークンの系列の生成には再帰や畳み込みを用いないことで計算量を抑えながら精度を向上させている特徴をもつTransformerを用いている．Transformerはエンコーダとデコーダからなる．

エンコーダでは，特徴抽出モジュールで生成された特徴マップを全結合層を用いて36×256のサイズに変換した系列 $(x_1, \ldots, x_{36})$を入力することでデコーダに入力するための系列$(z_1, \ldots, z_{36})$へ変換する．

デコーダでは，エンコーダで生成された系列$(z_1, \ldots, z_{36})$ および以前のデコーダの出力を用いて出力系列$(y_1, \ldots, y_m)$を生成する．出力系列の各要素が LaTeX形式のトークンに対応している．

3 表構造解析精度

Dengらの手法とFengらの手法に関して， ICDAR 2021[2]にて評価実験が行われた．評価実験では，43,138件のトレーニングデータ，800件の検証データおよび2,203件のテストデータの合計46,141件からなる表画像とそれに対応するLaTeXコードの組からなるデータセットが用意された．評価には生成されたトークンの系列と正解データの系列を比較して系列の全ての要素が一致している系列の割合を表すExact Match Accuracy（EM），系列の95%以上の要素が一致している系列の割合を表すExact Match Accuracy @95%（EM @95%），行数の一致精度を表すRow Accuracy（RA），列数の一致精度を表すColumn Accuracy（CA）の4つの評価指標があり，その評価実験の結果を表1に示す．表1より，CAの値は両手法で同じであり，それ以外の評価指標では僅かにFengらの手法の方がよかった．

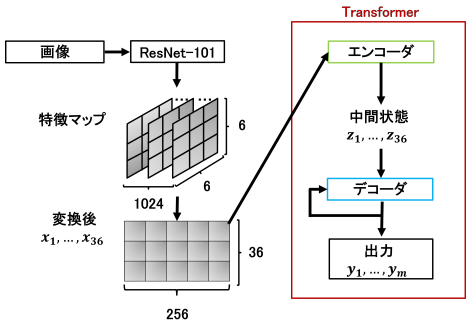

図3: Fengらの画像からトークンの系列を生成する手法の概要[5]

表1: 評価実験の結果[2]

Method	EM	EM @95%	RA	CA
Dengらの手法	0.66	0.79	0.92	**0.86**
Fengらの手法	**0.69**	**0.85**	**0.93**	**0.86**

4 おわりに

本稿では，表画像入力による表構造解析手法について調査した．調査した手法は，画像の特徴を抽出した後にLSTMまたはTransformerのエンコーダとデコーダを用いている．今後は，調査の結果優れていたFengらの手法で用いられた Transformerを用いた表画像入力による表構造解析に焦点を当てて研究を進める予定である．

参考文献

[1] 田上歩夢他， “表構造情報を利用した棒グラフの自動生成の一手法，” 第16回データ工学と情報マネジメントに関するフォーラム（DEIM 2024），T4-A-3-02，2024.

[2] Pratik Kayal *et al.*, “ICDAR 2021 Competition on Scientific Table Image Recognition to LaTeX,” In: *Proceedings of 16th International Conference on Document Analysis and Recognition – ICDAR 2021*, pp. 754-766, 2021.

[3] Yuntian Deng *et al.*, “Image-to-Markup Generation with Coarse-to-Fine Attention,” In: *Proceedings of the 34th International Conference on Machine Learning*, Volume 70, pp. 980–989, 2017.

[4] Sepp Hochreiter *et al.*, “LONG SHORT-TERM MEMORY,” *Neural Computation*, Volume 9, pp. 1735-1780, 1997.

[5] Xinjie Feng *et al.*, “Scene text recognition via transformer,” arXiv preprint arXiv:2003.08077, 2020.

[6] Kaiming He *et al.*, “Deep residual learning for image recognition,” In: *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[7] Ashish Vaswani *et al.*, “Attention Is All You Need,” *Advances in Neural Information Processing Systems 30*, pp. 5998-6008, 2017.