DEIM2021 D33-2

視覚化意図を考慮した データの効果的な視覚化方法の推定

丸田　敦貴† 　加藤　誠††

† 筑波大学 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1711567@s.tsukuba.ac.jp, ††mpkato@slis.tsukuba.ac.jp

あらまし　本研究では，双方向アテンションモデルを用いた自動視覚化システムを提案する．既存の自動視覚化推薦システムはデータの統計情報のみを使って視覚化方法の推定を行なってきたが，我々は「日本の人口の推移」などといったような視覚化意図を考慮してより適切な視覚化を推定する手法を提案する．我々は表形式データと視覚化のペアのデータセットを新たに構築し，実験を行なった結果，双方向アテンションを使ったモデルがベースラインを上回り，我々の提案するモデルの中で最も良い性能を示した．

キーワード　視覚化，表形式データ，双方向アテンション

1　はじめに

データの視覚化はデータの内容を伝えるのに効果的である．しかしながら，データの効果的な視覚化には専門的な知識 (あるデータに対して効果的なグラフは何か，視覚化ツールの使い方) やデータの精査が必要であり，特にデータの扱いに慣れていないエンドユーザにとっては大きな労力となり得る．自動的に視覚化を推薦するシステムがあればデータ視覚化の専門的な知識やデータの精査なしで効果的な視覚化を行うことができる．これに加えて，大量のデータの中から必要なデータを探さなければならない場合においては，もし適切に視覚化された中からデータを探すのであればデータの探索は容易であるが，視覚化がされていない場合，表形式データのような数値の羅列から目的のデータを探索するのは困難である．

これまでも，自動視覚化に関する研究は行われてきており，多くの研究は機械学習を用いた方法を採用している [1] [2] [3] [4]．この方法では，表形式データの統計情報，例えば行の数や列の値の分散を特徴として学習を行い，各データに適した視覚化種類（例えば，円グラフや棒グラフ，折れ線グラフなど）を推定している．しかしながら，既存の自動視覚化の手法には 2 つの問題がある．1 つ目に視覚化種類を推定するときユーザがどのような視覚化を行いたいか，という視覚化意図を考慮していない．適切な視覚化種類は表形式データだけからは一意に定まらず，表形式データの中のどの部分を視覚化するか，また，どのようなデータの傾向を表現するか，といった視覚化の意図に大きく左右される．例えば「日本の人口の推移」という視覚化意図が与えられたとき,「推移」という単語から折れ線グラフを使うのが適切だと考えられる，また「日本の人口の大学生の割合」という視覚化意図が与えられたとき,「割合」という単語から円グラフを使うのが適切だと考えられる．このように自動視覚化を行う際には表形式データのみならず，視覚化意図も考慮すればより効果的な視覚化を推薦することができると考えられる．2 つ目に，既存の研究では入力データとして表形式データの中の視覚化に用いる列，すなわち視覚化列しか用いていない．この問題設定では，オープンデータから視覚化を作る時に，データの精査をユーザが行う必要があり，自動視覚化を行うことはできない．そこで，視覚化意図を用いて表形式データの中から関係のある列を推測することで，ユーザが介入することなく適切な視覚化種類を推測することができる．

本論文では自動視覚化において,「日本の人口の推移」などといったような視覚化意図をデータと同時に考慮し，より適切な視覚化種類を推定する方法について提案する．より具体的には，まず表形式データの各列と視覚化意図の各単語をベクトル化する．次に各列と各単語の類似度をそれぞれ計算し，双方向アテンションを用いたモデル [5] を用いて重要度を加味した各列ベクトルと各単語ベクトルを得る．アテンションとはデータの中の重要な部分に着目して，その部分の情報をより多く使うことで，精度を高めるという手法である．双方向アテンションとは 2 つの入力データに対してアテンションを用いたもので，2 つの入力データからそれぞれデータの重要な部分をもう一方のデータから推定するという手法である．本研究は入力データとして視覚化意図と表形式データの 2 種類を用いているため，双方向アテンションを用いてそれぞれのデータの重要な部分を推定することが効果的だと考えられる．この手法を用いて重要度を加味した列，つまり視覚化に用いる可能性が高い列を推定するのと同時に，重要度を加味した視覚化意図の単語，つまり視覚化種類の推測に関係ない単語を除いた単語を推定するようなモデルを用いた．そして得られた重要度を加味した列ベクトルと単語ベクトルを結合し，そのベクトルを多層パーセプトロンで学習し，視覚化種類の予測を行う．さらに，追実験として視覚化種類と視覚化列の両方の推定も行った．視覚化列の予測の手法としては，アテンションの結果を用いて各列が視覚化に用いられているのかどうか二値分類を用いて予測を行った．

本研究の設定として表形式データと視覚化がペアになっているデータセットが必要であるが，それに適したデータセットが入手できなかったので，新たにデータセットを作成した．Tableau Public[1]という視覚化されたデータを共有している Web サイトから 183,427 の表形式データと視覚化のペアを収集した．これらのデータに対して前処理を行い，提案モデルの訓練を行なった．視覚化種類の予測の評価には適合率，再現率，F 値を用いた．視覚化列の予測の評価には R 精度と nDCG [6] を用いた．視覚化種類の予測結果は双方向アテンションを用いたモデルの結果がベースラインを大きく上回り，提案モデルの中でも双方向アテンションを用いたモデルが最も高い性能を示した．また，視覚化意図と表形式データの両方のデータを使っている時の方が性能が向上しており，視覚化種類の予測に視覚化意図を用いることの有用性が示された．さらに，予測精度が高い視覚化種類やアテンションがどのように機能しているかを分析した．しかし，視覚化列の予測は提案モデルが適切ではなかったため，ベースラインとほぼ同じ結果であった．

本研究の貢献は以下の通りである: 1) 自動視覚化推薦において表形式データと視覚化意図を用いる新たな問題設定に取り組んだ, 2) Tableau のデータから表形式データと視覚化のペアを収集し，視覚化推薦における新たなデータセット構築した, 3) 視覚化意図を考慮した推薦において双方向アテンションを用いたモデルが効果的に作用することを示した．

本稿ではまず 2 節で関連する研究を説明し，本研究との違いを説明する．次に 3 節で提案手法の詳細を説明する，ここでは視覚化意図や表形式データのベクトル化の方法や双方向アテンションの詳細や出力方法，学習方法を説明する．次に 4 節で実験の詳細として，データセットの構築方法，データの前処理，結果を説明する．最後に 5 節で結論として本稿のまとめを行う．

2 関 連 研 究

本節では自動視覚化推薦に関する既存研究ついて述べる．これまでも自動視覚化推薦の研究が行われており，様々な手法が提案されているが大きく分けてルールベースと機械学習ベースの 2 つに分けられる．

ルールベースとは表形式データの統計情報を用いて事前に決められたルールにしたがって視覚化を行う方法である [7] [8] [9] [10] [11] [12] ルールの例として行の数が大きいと円グラフの可能性は低い，などが挙げられる．このアプローチではデータ視覚化の専門家の意見を参考にしてルールを作成する．そのため，ルールの範囲内であれば確実に専門家が作成したような効果的な視覚化を行うことが可能である．しかし，ルールベースはあまり柔軟ではないので，列や行の数が大きいデータを入力すると効果を発揮することができない．また，この手法は入力データとして視覚化に用いられた列のみを用いているため，本研究の設定である視覚化に用いられていない列を含んだ表形式データを入力とする場合，どのようにして適用するかが明らかでない．さらに，ルールベースを適用するとなると Multipolygon chart のような既存手法で用いられなかった視覚化種類について新たなルールを定義しなければならない．新たなルールを定義するには専門家との協議が必要になり，高いコストがかかってしまうため，ルールベースを用いるのは困難であると考えられる．

機械学習ベースとは表形式データの統計情報 (列の値の分散や平均など) を用いて，機械学習の様々な手法を適用して視覚化を推薦する手法である [1] [2] [3] [4]．このアプローチは事前に定義されたルールを用いないため学習するデータの量と質によって訓練されたモデルの性能に差が出るが，より柔軟な出力を可能とする．Dibia と Demiralp は視覚化を機械翻訳として扱い，宣言的言語を用いて視覚化を行う手法を提案している [1]．この手法は表形式データを json 形式に変換しエンコーダーデコーダーモデルを用いて宣言的言語に変換することで視覚化を行う．Luo らはルールベースと機械学習ベースの両方を使って視覚化推薦を行う手法を提案している [3]．この手法は表形式データの統計情報を用いた機械学習で得た値と専門家によって作られたルールベースで得られた値の合計をランキングにして学習を行い，上位数件の視覚化推薦を行うものである．これらの手法 [1] [3] は我々の実験設定と比べて入力と出力の形式が異なっているため，我々の手法に適用することは困難である．Hu らの研究 [2] は表形式データから統計的な特徴を抽出し，その特徴を使って機械学習を行うものである．この研究では各列から抽出した特徴と，列と列のペアから抽出した特徴を集計したものを特徴として用いて，ニューラルネットワークを学習させ，視覚化種類と軸を予測している．この研究では表形式データの各列から抽出した特徴を用いて機械学習を行っている．我々の研究も同じように表形式データの各列をベクトル化したものを用いるため，この研究とは密接な関係がある．この研究 [2] の表形式データから特徴を抽出する方法を参考にして本実験を行う．

本節では関連研究について説明したが，我々の研究と既存の研究には主に 2 つの違いがある．1 つ目に，我々の研究では入力データとして表形式データだけでなく，視覚化意図も用いている．表形式のデータは複数の数値や文字の列で構成されているのに対して，視覚化意図は文字列から構成されている．既存の手法ではこれらの異なる種類のデータを用いて視覚化の推薦を行うことは困難である．2 つ目に，既存の研究では入力データとして視覚化に用いる数列しか与えられていないのに対して，本研究では表形式データを入力して，そこから使う列を選択的に使用しなくてはならない．そういった観点から既存の手法には限界がある．

3 提 案 手 法

本節では表形式データと視覚化意図から視覚化種類と視覚化列を予測するための提案モデルについて説明する．本研究の問題定義は視覚化意図と表形式データの 2 つの入力データがあったときに，事前に選ばれたいくつかの視覚化種類の中から適切なものを予測することと，表形式データの各列が視覚化列であるかどうかを予測することである．図 1 は本研究の提案手法の

1：https://public.tableau.com/s/

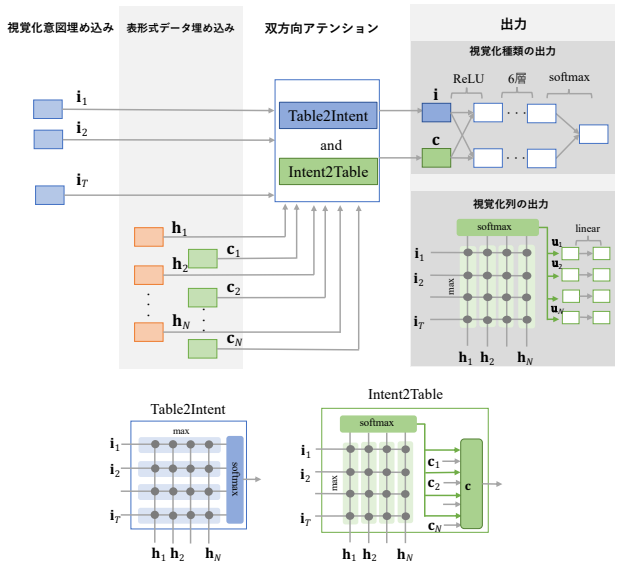

図 1　提案手法の大まかな流れ

大まかな流れを示している．我々の提案モデルは主に 4 つのコンポーネントから構成されている．(1) 視覚化埋め込み，(2) 表形式データ埋め込み，(3) 双方向アテンション，(4) 出力である．提案手法の中で重要なアイデアは双方向アテンション [5] を用いて視覚化意図から表形式データの重要な部分を，表形式データから視覚化意図の重要な部分をそれぞれ推定する 3 つ目のコンポーネントである．

3.1　視覚化意図埋め込み

1 つ目のコンポーネントとして視覚化意図の埋め込みについて説明する．まず，得られた視覚化意図を単語ごとに分割する．そして T 個の単語からなる視覚化意図 $X = \{x_1, x_2 \cdots, x_T\}$ が与えられたとき，任意の単語 x_t は One-hot 表現を用いて $\mathbf{v}_t$ と表す．ベクトル $\mathbf{v}_t$ は全語彙の集合 V を用いて作られる $|V|$ 次元のベクトルで，単語 s_t のインデックスに対応する次元は 1 それ以外を 0 とするベクトルである．また，v_t は単語埋め込み行列 $\mathbf{E}_w \in \mathbb{R}^{d_e \times |V|}$ を用いて単語ベクトル $\mathbf{i}_t = \mathbf{E}_w \cdot \mathbf{v}_t$ とする．d_e は単語ベクトルの次元数を表している．このようにして我々は視覚化意図ベクトル $\mathbf{i}_t \in \mathbb{R}^{d_e}$ を得る．

3.2　表形式データ埋め込み

2 つ目のコンポーネントとして表形式データの埋め込みについて説明する．表形式データはヘッダーと列の値の部分で別々の埋め込み方法を行う．ヘッダーとは各列の 1 番上に位置するセルの値のことで，ヘッダーはその列の概要を単語で表現していることが多い．そのため重要な列を推定するときにヘッダーの情報は有用であると考えられる．また，既存の研究で視覚化列の統計情報から視覚化種類を推定する方法が存在していたので，重要な列を推定するためのヘッダーの埋め込みと，視覚化種類を推定するための列の統計情報を用いた埋め込みの 2 種類の埋め込みを行う．まずヘッダーの埋め込みを説明する．N 列の表形式データが与えられたとき，j 列目のヘッダーベクトルを $\mathbf{h}_j \in \mathbb{R}^{d_e}$ と表す．j 列目のヘッダーを単語に分割して，M_j 個の単語を得たとき，それぞれの単語を視覚化意図と同じように埋め込み，ヘッダーの全単語ベクトルの平均値 $\mathbf{h}_j = \frac{1}{M_j}\sum_{u=1}^{M_j} \mathbf{i}_u$ をヘッダーベクトルとして用いる．ヘッダーの全単語の平均値を取ることで，ヘッダーの全体の意味を表現できる．次に，列の統計情報を用いた埋め込みを説明する．j 列目の統計情報ベクトル $\mathbf{c}_j$ は視覚化推薦の既存の研究 [2] で用いられた 1 つの列から得られる $d_c = 76$ 個の特徴を用いて埋め込みを行う．この特徴はカテゴリカル値と数値で構成されており，例えば，列の値の合計や平均，値が文字列かどうか，といった特徴がある．この特徴は関連研究 [2] において，有効であると示された特徴であるため，この特徴を用いる．

3.3　双方向アテンション

次に 3 つ目のコンポーネントの双方向アテンションについて説明する．アテンションとはデータの中の重要な部分に着目して，その部分の情報をより多く使うことで，精度を高めるという手法である．双方向アテンションとは 2 つの入力データに対してアテンションを用いたもので，それぞれの重要な部分をもう一方のデータから推定する手法である．本研究で双方向アテンションを用いる理由は視覚化意図と表形式データの 2 つの入力を効果的に利用することができるためである．本実験では視覚化に用いない列が含まれている表形式データを入力しているため，視覚化において重要な列を推定する必要がある．双方向アテンションを用いれば，表形式データの重要な列は予測可能であり，視覚化意図の中で表形式データと親和性の高い単語を予測することで視覚化に直接関係のない単語の情報を用いずに学習することができる．双方向アテンションではヘッダーベクトルと視覚化意図ベクトルの類似度を求める部分と表形式データから視覚化意図の重要な単語を推定するという部分 (Table2Intent) と視覚化意図から表形式データの重要な部分を推定するという部分 (Intent2Table) の 3 つに分かれている．

まず，各ヘッダーベクトルと視覚化意図ベクトルの類似度を計算する．t 番目の視覚化意図ベクトル $\mathbf{i}_t$ と，j 列目のヘッダーベクトル $\mathbf{h}_j$ を用いて，t 番目の視覚化意図の単語と j 列目のヘッダーの類似度 $s_{tj} \in \mathbb{R}$ は以下の式 1 のように得られる：

$$s_{tj} = \alpha(\mathbf{i}_t, \mathbf{h}_j) \tag{1}$$

α は学習可能な重みを持った関数であり，$\alpha(\mathbf{i}, \mathbf{h}) = \mathbf{w}_{(s)}^{\mathsf{T}}[\mathbf{i}; \mathbf{h}; \mathbf{i} \circ \mathbf{h}]$ と表すことができる．$\mathbf{w}_{(s)}$ は学習可能な重みベクトルで，$\circ$ はアダマール積を表している．$[;]$ はベクトル同士の結合を表している．

次に，表形式データから視覚化意図の重要な部分を推定する部分 (Table2Intent) を説明する．ここでは，ある視覚化意図の単語から得られるアテンションの重み a_t を求め，視覚化意図ベクトル $\mathbf{i}_t$ と掛け合わせる．まず，類似度 s_{tj} を用いてアテンションの重み a_t を以下の式 2 のように求める:

$$a_t = \mathrm{softmax}(\max_j(s_{tj})) \tag{2}$$

$\max_j$ は視覚化意図の単語ごとに最大値をとる関数である．こうして得られた重み a を用いて重要度を加味した視覚化意図ベ

クトル $\mathbf{i} \in \mathbb{R}^{d_e}$ を以下の式 3 のようにして求める:

$$\mathbf{i} = \sum_{t=1}^{T} a_t \mathbf{i}_t \tag{3}$$

ある視覚化意図ベクトルとヘッダーベクトルの類似度が小さい値のときアテンションの重み a_t は 0 に近づき，視覚化意図ベクトル $\mathbf{i}$ も小さい値となるため，表形式データとの親和性が低い視覚化意図の情報は用いないという構造になっている．

次に，視覚化意図から表形式データの重要な部分を推定する部分 (Intent2Table) を説明する．列ごとに得られるアテンションの重み b_j を求め，列の統計情報ベクトル $\mathbf{c}_j$ と掛け合わせる．Table2Intent と同じように，アテンションの重みを以下の式 4 のように求める:

$$b_j = \mathrm{softmax}(\max_t(s_{tj})) \tag{4}$$

$\max_t$ は列ごとに最大値をとる関数である．こうして得られた重みを用いて重要度を加味した列ベクトル $\mathbf{c} \in \mathbb{R}^{d_c}$ を以下の式 5 のようにして求める:

$$\mathbf{c} = \sum_{j=1}^{J} b_j \mathbf{c}_j \tag{5}$$

アテンションによって得られた 2 つのベクトルからアテンションベクトル $\mathbf{x} \in \mathbb{R}^{d_e+d_c}$ を以下の式 6 のように定義する:

$$\mathbf{x} = \mathbf{i}; \mathbf{c} \tag{6}$$

ベクトル $\mathbf{x}$ を用いて多層パーセプトロンの学習を行う．

双方向アテンションで行われていることの例として「日本の人口の推移」という視覚化意図と「人口増加率」,「年」,「平均年齢」がヘッダーであるような 3 列の表形式データが入力されたときに，Table2Intent では視覚化意図の単語の中で「人口」という単語に近い意味のベクトルを得る．なぜなら「人口」という単語は表形式データのヘッダーの「人口増加率」と類似度が高い単語だからである．Intent2Table でも同じように表形式データの列の中で「人口増加率」という列の統計情報ベクトルに高い重みがついたベクトル得る．このようにして双方向アテンションでは類似度から視覚化意図と表形式データの重要な部分を推定している．

3.4 出　　力

最後のコンポーネントとして出力を説明する．本研究では視覚化種類の予測と視覚化列の予測の 2 つの出力があるので，それぞれについて説明する．

3.4.1 視覚化種類の予測出力

アテンションで得られたベクトル $\mathbf{x}$ を多層パーセプトロンに入力し学習させ，事前に決められたいくつかの視覚化種類の中で最も適切な種類を 1 つ予測する．まずアテンションベクトル $\mathbf{x}$ を式 7 のような非線形変換を行う全結合層に入力する，この層はいくつかの層で構成されている:

$$\mathbf{g} = \mathrm{ReLU}(\mathbf{w}_g \mathbf{x} + b_g) \tag{7}$$

$\mathbf{w}_g \in \mathbb{R}^{T+J}$，$b_g \in \mathbb{R}$ は全結合層のパラメータで ReLU は Rectified Linear Unit である．$\mathbf{g}$ は出力ベクトルであり，予測ラベル $\mathrm{p_t} \in \mathbb{R}$ は以下の式 8 のように定義する:

$$\mathrm{p_t} = \arg\max(\mathrm{softmax}(\mathbf{g})) \tag{8}$$

3.4.2 視覚化列の予測出力

視覚化意図の予測では列ごとに得られるアテンションの重み b_j を入力としてその列が視覚化列かそうでないかを予測する．重み b_j を 1 度式 9 のような線形変換を行って予測を行った：

$$u_j = w_{u_j} b_j + b_{u_j} \tag{9}$$

$w_{u_j} \in \mathbb{R}$，$b_{u_j} \in \mathbb{R}$ は全結合層のパラメータで列ごとの出力 u_j を学習に用いた．

3.4.3 学　　習

本研究は出力データが 2 つあり，視覚化種類を予測するモデルと視覚化列を予測するモデルを両方学習する必要があるのでマルチタスク学習を行う．この学習方法は両方のモデルを学習していく中で 2 つのモデルで共有している部分を作り，各モデルがそれぞれ 1 回学習する間に共有している部分を 2 回学習させるという手法である．損失関数は視覚化種類の予測モデルの学習では交差エントロピー誤差 L_{ce} を用いて，視覚化列の予測では二値交差エントロピー誤差 L_{bce} を用いる．損失関数とは出力された予測値と正解ラベルの差，つまり損失を計算する関数であり，損失が小さくなることで学習が進む．これら 2 つの損失関数をまとめた損失関数 L を以下の式 10 のように定義する:

$$L = \beta L_{ce} + (1 - \beta) L_{bce} \tag{10}$$

β はどちらの損失関数を重視するかというパラメータであり，視覚化種類の予測のみを行う場合は $\beta = 1$ とする．パラメータの最適化には Adam [13] を用いる．Adam は確率的勾配降下法を拡張したもので，他の最適化方法と比べ優れていることが示されている．

4 実　　験

本節では実験の詳細としてデータセットの構築方法，データの前処理，実験設定，実験結果を説明する．

4.1 データセット

我々の研究では学習に必要なデータとして視覚化意図，表形式データ，視覚化種類 (円グラフや棒グラフなど)，視覚化列の 4 つを取得する必要がある．しかしながら，既存の研究 [1] [2] [3] [4] では視覚化列のみを入力データとしており，視覚化に用いない列を入力していた研究は行われていなかった．視覚化列のみを入力する実験方法の場合，オープンデータから表形式データを取得し，視覚化を作る際にデータを理解し，その中で視覚化列を選ばなくてはならない．我々の実験はデータを理解せずとも自動で視覚化を行うことを目標としているので，視覚化に用いない列を含んだデータが必要であった．そこで視覚化推薦のためのデータセットを新たに構築した．

我々は視覚化と表形式データを Tableau Public から収集した．Tableau Public は視覚化されたデータを共有している Web サイトである．Tableau Public から視覚化のタイトル，表形式データ，視覚化種類，視覚化列をクローリングした．クローリングとは Web 上にある特定の情報のみを抽出することである．タイトルは Table Public のブラウザ上で視覚化と共に記載されている文章を取得した．視覚化のタイトルは視覚化を文章で表現したものであるため，本研究ではタイトルを視覚化意図として用いた．Tableau Public では複数の視覚化をまとめたダッシュボードが存在する．ダッシュボードの中のそれぞれの視覚化にもタイトルはついているが，それでは「gender」や「year」といった抽象的すぎるタイトルが多かったため，ダッシュボードのタイトルとそれぞれの視覚化のタイトルを結合したものを視覚化意図として用いた．Tableau Public は何らかの検索ワードを入力しないと視覚化にアクセスすることができないので，検索ワードを決める必要があった．また，Tableau Public では検索結果上位 10,000 件しか取得できなかった．そこで，NLTK[2]の Reuters corpus に出現する各単語を出現頻度順に並べて，Tableau Public の視覚化検索結果が約 10,000 件の単語から頻度の少ない単語に向かって 1 語ずつ検索した．これはデータの重複を避けて効率的にクローリングを行うためである．

4.2 前 処 理

視覚化意図と列の統計情報のベクトルに対して行った前処理について説明する．まず，視覚化意図の前処理として，単語ごとに分割し，括弧や句点のような記号を取り除き，単語数が 3 語以下のものは取り除いた．この処理を行う理由は単語数が少ない視覚化意図は略語や人間でも理解できないタイトルであることが多いためである．

次に列の統計情報の前処理について説明する．列の統計情報ベクトルは，値があまりにも大きすぎる場合，データを読み込むときに無限大という値に変換されてしまい，学習不可能となるため無限大の値を各特徴の無限大の値を除いた最大値に置き換えた．次に，99 パーセンタイルを超える値，または 1 パーセンタイルを下回る値をカットオフした．カットオフとは，ある値を境にその値以上あるいはその値以下の値をすべてある値に置き換える処理のことである．この処理によって極端に大きな値や小さい値を除くことができるため効果的な学習を行うことができる．次に欠損値の対策として数値の欠損値は平均値で補い，カテゴリカル値の欠損値は最頻値で補った．この処理を行う理由は，特徴の中に無限大の値と同様に欠損値が存在すると学習不可能となるためである．最後に特徴ごとに標準化を行った．標準化とはデータ平均を 0 に，分散を 1 にする処理のことで標準化を行うことで特徴ごとの値の大きさの差をなくすことができる．標準化によって，特徴ごとのデータの傾向 (平均値など) を一様にすることで，効果的な学習を行うことができる．

前処理の結果，データ数が 183,427 となり，訓練データ (148,575 データ)，検証データ (16,509 データ)，テストデータ (18,343) の 3 つに分割した．データを分割したのは，モデルを学習させるための訓練データとパラメータのチューニングをするための検証データと学習モデルの性能を図るためのテストデータに分けることで，1 度学習したデータがモデルの性能を測るテストに出てこないようにするためである．ハイパーパラメータとはモデルの性能を少し変化させることができる数値のことで，これは入力するデータによって最適な値が変化するため，最適なパラメータを得るために検証データが用いられている．また，視覚化意図の平均単語数は 7.9 語で，表形式データの平均列数は 16.7 列であった．

4.3 実 験 設 定

単語埋め込みには GLoVe [14] を用いた事前学習済みの分散表現モデル用いた．我々が用いたのは Wikipedia2014 と Gigaword5[3]の文章から学習されたものである．また，単語埋め込みの次元数は $d_e = 100$ に設定した．埋め込みができなかった単語 (事前学習済みのモデルに含まれていない単語，英語ではない単語や略語など) が 1 つの視覚化意図の単語数の半数を超える場合，意図が薄れてしまうのでその視覚化意図は取り除いた．単語数は 12 語の上限を設けた．表形式データは 30 列を列数の上限としたため，視覚化列予測の評価に用いる nDCG の測定長は 30 とした．本研究では 8 種類の視覚化種類 (Area, Bar, Circle, Line, MultiPolygon, Pie, Shape, Square) を予測に用いた．予測に用いる視覚化種類を 8 種類にした理由は，他の視覚化種類のデータ数が極端に少なかったためである．検証データを用いた結果から多層パーセプトロンの層の数は 6 層とした．効率の良い学習を行うためにバックプロパゲーション [15] を用いた．また，過学習を防ぐために，多層パーセプトロンのドロップアウト率 [16] は 0.2 とした．

結果の比較対象として機械学習ベースの関連研究の手法を用いた [2]．この手法では 1 つの表形式データから 912 個の特徴を抽出して視覚化種類を予測しているため，この特徴を用いたモデルを提案手法のベースラインとして用いる．他の機械学習ベースの関連研究は本実験の設定に対して入力と出力の形式が異なっているため，比較することができなかった [1] [3] [4]．提案手法と比較するベースラインモデルとして以下のものを設定した:

- **Naive Bayes** :ナイーブベイズ [17]
- **K-Nearest Neighbor** :K-近傍法 [18]
- **Logistic Regression** :ロジスティック回帰 [19]
- **Random Forest** :ランダムフォレスト [20]

検証データを用いてそれぞれのモデルのハイパーパラメータをチューニングした．

4.4 実 験 結 果

表 1 はベースライン手法と我々の提案手法の性能の差を 3 つの評価指標である適合率，再現率，F 値を用いて示している．ベースラインモデルも高い値を示しているが，視覚化意図と表

2：https://www.nltk.org/

3：https://catalog.ldc.upenn.edu/LDC2011T07

表 1　提案手法と比較手法の結果

カテゴリ	モデル	適合率	再現率	F 値
ベースライン [2]	ナイーブベイズ	0.225	0.144	0.058
	K-近傍法	0.479	0.432	0.449
	ロジスティック回帰	0.342	0.208	0.204
	ランダムフォレスト	0.478	0.439	0.454
提案手法 (アテンションなし)	意図	0.503	0.456	0.471
	表	0.442	0.385	0.404
	意図 & 表	0.548	0.518	0.530
提案手法 (アテンションあり)	意図 + アテンション	0.519	0.476	0.490
	表 + アテンション	0.458	0.414	0.430
	意図 & 表 + アテンション	**0.561**	**0.544**	**0.551**

形式データの両方を使ったモデルではないので提案手法と比べることができない．そのため双方向アテンションを用いていないモデルと比較を行う．“アテンションなし” は提案手法を単純なモデルに変更したもので重要度を加味した視覚化意図ベクトル $\mathbf{i}$ を視覚化意図ベクトル $\mathbf{i}_t$ の平均をとった値に変更し，重要度を加味した統計情報ベクトル $\mathbf{c}$ を表形式データの統計情報ベクトル $\mathbf{c}_j$ の平均をとった値に変更したものである．“意図&表” は視覚化意図ベクトル $\mathbf{i}$ と表形式データの統計情報ベクトル $\mathbf{c}$ を組み合わせたベクトル $\mathbf{x}$ を多層パーセプトロンの入力に用いた結果である．“意図” と “表” はそれぞれ多層パーセプトロンに入力するベクトルを $\mathbf{i}$ のみと $\mathbf{c}$ のみにした結果である．我々の提案した双方向アテンションを用いたモデルがすべての指標においてベースラインを上回り，他の単純なモデルの提案手法の中でも最も高い性能を示していることがわかる．また，双方向アテンションを用いていないモデルと比較してもそれぞれのモデルで性能が上がっていることがわかる．この結果から，本研究の提案モデルである双方向アテンションが効果的に働いたと考えられる．また視覚化意図のみと統計情報のみを用いた時の結果を比較すると，視覚化意図の方が高い性能を示しており，表の統計情報よりも視覚化意図の情報の方が視覚化意図の予測に効果があることが示された．また，視覚化意図と表の情報を組み合わせて使うとより高い性能を示した．この結果から，視覚化種類の予測に視覚化意図を用いるのは効果的だったと言える．このモデル間の結果に有意な差があるかを検証するために，ボンフェローニ補正を用いた並べ替え検定を行った．サンプル数は 10,000 データで，最も高い性能を示したモデルとそれ以外の各モデルの 9 ペアに対して検定を行ったところ，$\alpha = 0.01$ で有意な結果が得られた．並べ替え検定とは 2 つの母集団には差がないという帰無仮説を棄却するもので，2 つの母集団から無作為に抽出した 2 つの標本を無作為に 2 つのグループに割り当て，グループ間で平均値のような統計量が異なれば 2 つの母集団の間には有意な差があるという検定方法である．しかし 3 つ以上の比較を行う場合に検定の多重性が生まれるため，ボンフェローニ補正を行った．検定の多重性とは例えば，3 回検定を行いすべて 5%の有意水準を下回る確率を考えたときに，$1-(1-0.05)*(1-0.05)*(1-0.05) = 0.143$ となり有意である確率が増えてしまう，このように複数回検定を行ったときに有意である確率が上がってしまうことを検定の多重性という．

表 2　視覚化意図ごとの提案手法 (意図&データ + BiDA) の結果

視覚化方法	# データ数	適合率	再現率	F 値
Area	978	0.555	0.443	0.493
Bar	4,506	0.551	0.600	0.575
Circle	3,464	0.517	0.529	0.523
Line	2,485	0.513	0.493	0.503
Multipolygon	1,399	0.625	0.674	0.649
Pie	2,075	0.595	0.611	0.603
Shape	1,877	0.558	0.541	0.550
Square	1,559	0.572	0.464	0.512

表 2 は視覚化種類ごとの結果を示しており，表 3 は視覚化種類ごとのアテンションの値が高かった単語，つまり予測において有効であった単語を示している．視覚化意図の中で最も高い予測精度を示したのは Multipolygon chart である．Multipolygon chart のアテンションの値が高く示された単語を見ると *geographical* や *zip* など土地に関係する単語が含まれていた．Multipolygon chart のみが地図に関係する視覚化であるため，高い予測精度を示したと考えられる．次に予測の精度が高かったのは Pie chart である．Pie chart のアテンションの値が高かった単語に *pie* や *donut* など視覚化の形そのものに関する単語が多く含まれていたため，高い予測精度を示したと考えられる．これらの高い予測精度を示した視覚化は独特な形をしていたり，他の視覚化にはない特徴を持っていたため予測の精度が高く示されたと考えられる．予測の精度が低く示されたのは Area chart と Line chart であり，アテンションの値が高く示された単語を見ると，両方に *trend* や *area* が含まれていた．これらの視覚化種類の予測の精度が低かった理由として，Area chart と Line chart は両方とも点を線で繋いだ視覚化のため形が似ており，視覚化意図にも似た表現が多く含まれていたため，2 つの視覚化を分類できなかったためだと考えられる．

図 2 はアテンションの成功例を図示したものである．この図は正解ラベルが Line chart で正しく予測することができた時の視覚化意図と表形式データのヘッダーの類似度をヒートマッ

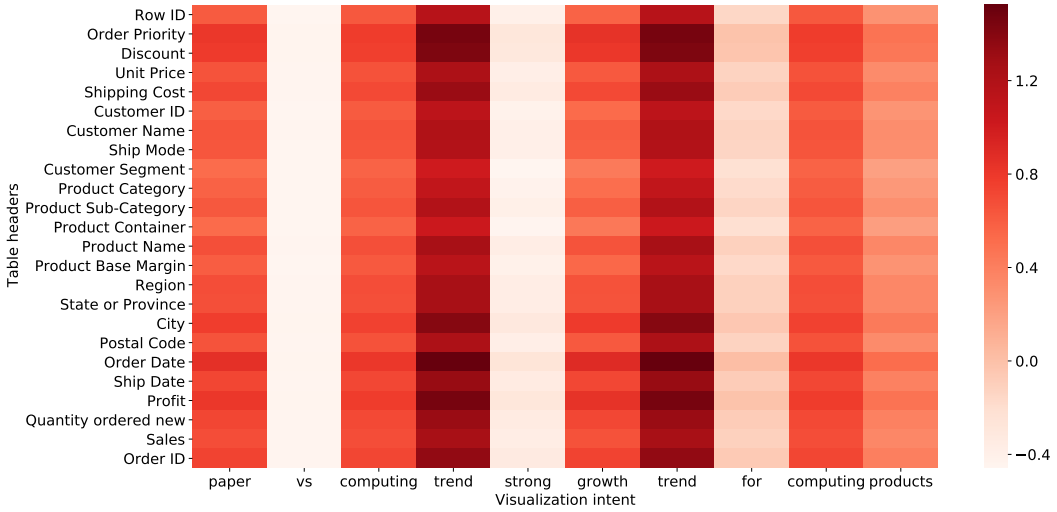

図 2　類似度行列の視覚化．正解データが Line で提案モデルは正しく予測した．視覚化意図の中で trend という単語が最もアテンションの値が高い．

プで表している．予測のときに大きな重みをつけた場所，つまり重要な場所だと予測した部分は濃い色で，そうでない場所は薄い色で表している．図を見ると視覚化意図の単語ごとに大きく色が変化しており，類似度はヘッダーにあまり影響を受けず，視覚化意図に大きな影響を受けていると考えられる．視覚化意図の単語では予測に効果的である *trend* のアテンションの値が高く示されている．また *vs* や *for*，*strong* と言った視覚化種類の予測に関係のない単語はアテンンションの値が低く示されている．そのため，視覚化意図の情報を選択的に取得していると言える．表形式データのヘッダーでは，視覚化列は *Order Date* と *Sales*，*Product Sub-Category*，*State or Province* であったが，*Order Date* の部分は高い値を示しており，1 列の予測には成功した．

図 4 は視覚化種類と視覚化列の予測を同時に行なった結果を示している．この実験は本実験の後に追実験として行ったため，視覚化列の予測の前処理でデータ数が減少し 159,494 データとなった．データ数が減った理由は，実験設定で使用する列は 30 列という上限を決めているので，31 列目以降に視覚化列があった場合にそのデータを使用しなかったためである．視覚化列の予測ではその列が使用される確率が出力されるため確率の高さでランキングを作り，ランキングの評価指標として R 精度と nDCG [6] 用いた．R 精度とは視覚化列の数だけランキングの上位から判定したとき，視覚化列が含まれる割合である．“モデル” 列のランダムは，視覚化列予測ではランダムなランキングを出力し，視覚化種類の予測ではランダムな視覚化種類を出力したものを評価した結果である．視覚化種類の予測では学習のパラメータ β によって精度の差が生まれたが，視覚化列の予測ではどのモデルとも精度の差がなく，ランダムベースラインの結果とほぼ変化はなかった．この結果から，今回の視覚化列を予測する提案モデルは効果的ではなかったと考えられる．さらに $\beta = 0.5$，つまり視覚化列の予測モデルと視覚化種類の予測モデルを同じ程度に学習させたとき，視覚化種類の予測精度が低い値を示した．視覚化種類の予測精度が低かった理由として，視覚化列の予測モデルと視覚化種類の予測モデルの両方のモデルを効果的に学習することができなかったためだと考えられる．この結果から，両方の予測結果を出力するのであればマルチタスク学習を行うのではなく，それぞれのモデルを独立して学習させた方が良いと考えられる．

5 ま と め

本研究では視覚化意図と表形式データを組み合わせた視覚化推薦システムを提案した．我々の提案した双方向アテンションを用いたモデルは視覚化意図の重要な部分と表形式データの重要な部分を予測するモデルである．我々は表形式データと視覚化の新たなデータセットを構築し，実験を行った結果，視覚化種類の予測で双方向アテンションを用いたモデルがベースラインを上回り，提案モデルの中でも最も高い性能を示した．しかし，視覚化列の予測ではベースラインとほぼ変わらない結果であったため，提案モデルをさらに改良する必要があった．本研究では視覚化種類と視覚化列の予測に取り組んだが，今後の課題として，今回取り組んだものだけでなく軸やレイアウトのような細かい視覚化の設定も予測することが挙げられる．

謝辞　本研究は JSPS 科研費 18H03244，18H03243，および，JST さきがけ JPMJPR1853 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Victor Dibia and Çağatay Demiralp. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE computer graphics and applications*, 39(5):33–46, 2019.

[2] Kevin Hu, Michiel A Bakker, Stephen Li, Tim Kraska, and César Hidalgo. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[3] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. Deepeye: Towards automatic data visualization. In *2018 IEEE 34th*

表 3　視覚化種類ごとのタイトルの中で最もアテンションの値が高くなった単語．“アテンション” 列の値はテストデータのなかの各単語のアテンションの値の平均である．

種類	単語	アテンション
Area	trend	0.19
	area	0.16
	sheet	0.15
	state	0.14
	chart	0.14
Bar	axis	0.26
	bar	0.25
	filter	0.22
	bars	0.22
	table	0.21
Circle	bubble	0.28
	donut	0.26
	map	0.24
	box	0.24
	across	0.20
Line	bar	0.26
	across	0.23
	trend	0.19
	trends	0.18
	area	0.17
Multipolygon	zip	0.25
	map	0.25
	across	0.24
	geographical	0.22
	median	0.16
Pie	pie	0.33
	map	0.26
	donut	0.24
	you	0.17
	country	0.15
Shape	map	0.26
	filter	0.25
	icon	0.21
	circle	0.21
	trend	0.20
Square	map	0.26
	table	0.23
	country	0.16
	sheet	0.15
	state	0.15

表 4　視覚化列と視覚化種類の予測の結果

	視覚化列予測		視覚化種類予測		
モデル	R 精度	nDCG@30	適合率	再現率	F 値
ランダム	0.127	0.452	0.15	0.12	0.13
$\beta = 0.0$	0.129	0.457	0.15	0.13	0.12
$\beta = 0.5$	0.125	0.455	0.51	0.51	0.50
$\beta = 1.0$	0.126	0.455	0.53	0.53	0.52

International Conference on Data Engineering (ICDE), pages 101–112. IEEE, 2018.

[4] Dominik Moritz, Chenglong Wang, Greg L Nelson, Halden Lin, Adam M Smith, Bill Howe, and Jeffrey Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics*, 25(1):438–448, 2018.

[5] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv.org e-Print archive*, 2016, 1611.01603. https://arxiv.org/pdf/1611.01603.pdf, (accessed 2020-12-19).

[6] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[7] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2015.

[8] Jock Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

[9] Steven F Roth, John Kolojejchick, Joe Mattis, and Jade Goldstein. Interactive graphic design using automatic presentation knowledge. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 112–117, 1994.

[10] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002.

[11] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007.

[12] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv.org e-Print archive*, 2016, 1611.01603. https://arxiv.org/pdf/1611.01603.pdf, (accessed 2020-12-19).

[14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[15] James L McClelland, David E Rumelhart, PDP Research Group, et al. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[17] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[18] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[19] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[20] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.