分散表現に基づく漫画の画像と台詞の対応識別

1 はじめに

近年，深層学習をはじめとする機械学習技術の発展を背景として，人工知能による小説やイラストの自動生成など計算機による創作物理解が注目されている．しかしながら，人の創作は高次の知的活動であるため，現段階では表層的な模倣以上の成果は得られていない．

人の創作物の理解に関する分野の中でも絵および文字から構成される漫画 (コミック) を対象とする研究が盛んになされてきた．これは，コミックを扱う研究が自然言語処理と画像処理の両方の側面を持つマルチモーダルなデータを対象とする挑戦的な分野であるためである．同分野では従来研究として自然言語情報を用いて漫画のコマの順序を識別する手法や，画像情報を用いて登場人物の位置を推定する手法などが報告されている．しかしながら，漫画の台詞におけるスパース性やマルチモーダルを扱うモデルの難しさなどから，画像と言語の両方の側面からのアプローチをする研究はほとんどされてこなかった．

そこで本研究では，これらの問題を解決するために Manga109 [1][2] および 4 コマ漫画ストーリーデータセット [3] を用いて，漫画の計算機によるマルチモーダル的な意味理解の研究を進める．本研究では漫画をマルチモーダル的に扱う研究への取り掛かりとして，4 コマ漫画におけるコマ画像および台詞の分散表現に基づく画像と台詞の対応識別のタスクに取り組む．

2 研究用コミックデータ

2.1 Manga109

Manga109 は漫画を扱う学術研究のために相澤，山崎研究室により作成されたコミックデータセットである． Manga109 は 104 種類のストーリー漫画と 5 種類の 4 コマ漫画によって構成されている．また，データセットの特徴としてはキャラクタの全身や顔，台詞等の座標情報が付与されていることや，登場人物の名前，発話内容等の情報がまとめられているなど，画像データセットとしても言語データセットとしても豊富なアノテーションがなされている点が挙げられる．本研究では Manga109 より 4 コマ漫画である 5 作品 (後述) を用いた．

2.2 4 コマ漫画ストーリーデータセット

本研究では Manga109 に加え，4 コマ漫画ストーリーデータセットを用いる．このデータセットは人工知能と創作物に関する研究発展のために研究者が一から開発に関わった世界初の研究用データセットであり，いくつかの特徴がある．まず，Manga109 のように市販の漫画を用いて作成されたデータセットとは異なり，4 コマ漫画ストーリーデータセットは本データセットのために同一のストーリーに対して幾人かの漫画家によって描き下ろされている．市販の漫画を用いた場合，著作権などの問題に加え，計算機上で扱うための著者の感性に関する情報が少なく，漫画の意味理解を目的とした研究には適さないという問題がある．例えば漫画に登場するキャラクタの感情が読者によってアノテートをされた際，アノテートされたラベルが漫画家の意図とは異なる可能性を否定できない．4 コマ漫画ストーリーデータセットは上述の問題点を解決するために作られたデータセットであり，漫画を扱う人工知能研究において大きな利点がある．

また，漫画は異なる作者によって描かれており，そのタッチを基にギャグタッチ，少女漫画タッチ，少年漫画タッチ，青年漫画タッチ，萌えタッチの 5 タッチに分類されている．本研究ではこれらのタッチのうち，少年漫画タッチを用いる．

3 要素技術

3.1 分散表現化モデル

3.1.1 自己符号化器

自己符号化器 (AutoEncoder: AE) とはニューラルネットワークの一種で，情報量を落とさずにより低次元の特徴表現を獲得するための次元圧縮手法である．AE は出力データが入力データと一致するように損失関数を定め，学習を進める．

3.1.2 畳み込み自己符号化器

畳み込み自己符号化器 (Convolutinal AutoEncoder: CAE) は 3.1.1 節で説明した AE に畳み込み層を追加したものである． CAE ではエンコーダに畳み込み層と MaxPooling 層を用いており，デコーダに畳み込み層と UpSampling 層を用いている． CAE は畳み込み層を有しているため，特に画像処理に対して有効である．

3.1.3 BERT

Bidirectional Encoder Representations from Transformers(BERT) [4] とは Transformer [5] を拡張した双方向のエンコーダモデルである．BERT の特徴は単語及び文の汎用的な分散表現を獲得できる点にあり，事前学習済みの BERT によって得られた分散表現により，Stanford Question Answering Dataset(SQuAD) における質疑応答タスクや MultiNLI のおける自然言語推論などの幅広い自然言語処理タスクにおいて最先端の結果が出されている．本研究では日本語 Wikipedia 全文 (約 1,800 万文) を用いて事前学習した BERT モデルを使用した．

3.2 LSTM

Long Short-Term Memory(LSTM) [6] はループ構造を持ち，過去の時系列情報を記憶する Recurrent Neural Network (RNN) を拡張して作成された機械学習モデルである．RNN には時系列が長くなることで勾配が消失してしまうという問題点があったが，LSTM ではゲートと呼ばれる仕組みによってこれを解消している．また，LSTM にデータ系列を正順と逆順で別々に入力し，中間層出力を連結するモデルを双方向 LSTM と呼ぶ．双方向 LSTM は未来方向と過去方向両方の情報を用いることができるため，より時系列データ解析に有効であるという特徴をもつ．

4 提案手法

本研究では漫画のコマ画像および台詞を分散表現化し，その対応識別タスクを解くことを目的とする．しかしながら，コマ画像と台詞の対応識別を直接的に解いたり，画像に合うような台詞を生成したりするタスクは現状では困難である．これは，1 つのコマ画像に対しストーリー性を崩さないような台詞の入れ替えというものは多数存在するため極めて自由度の高い問題なためである．また，1 つの台詞が与えられたときにキャラクタ，吹き出しおよび背景などをどのように配置するかというようなコマの構図も任意性が高いといえる．

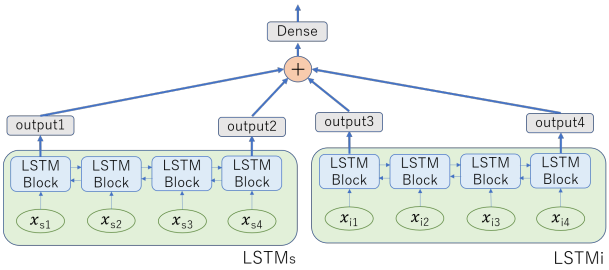

図 1: 提案双方向 LSTM モデル

よって本研究では，漫画のコマ画像と台詞を用いたマルチモーダル研究の取り掛かりとして，漫画の 4 コマ目の画像や台詞を入れ替えた時に入れ替えたことを検知できるかというタスクを設定した．4 コマ漫画には 4 コマ分のコマ画像と台詞がそれぞれ時系列データであるという特徴があるため，モデルとしては時系列情報を扱うのに優れている双方向 LSTM をベースとしたモデルを構築した．図 1 に本研究で使用した提案モデルを示す．提案モデルにおいては，画像と台詞でそれぞれ双方向 LSTM($\mathrm{LSTM_i}$, $\mathrm{LSTM_s}$) を定義し，その最終層を連結し，Multi Layer Perceptron(MLP) を用いて画像および台詞を入れ替えたかどうかいう識別タスクを解く．また，本研究では LSTM の入力にコマ画像および台詞の分散表現を用いる．画像の分散表現を得るために CAE を用い，台詞の分散表現を得るために事前学習済み BERT を用いた．

5 数値実験

本研究では，3 パターンのデータセットについてコマ画像と台詞の対応識別実験をする．本実験では，最適化手法に Adam，損失関数には交差エントロピーを用い，100 エポック分学習した．訓練と評価画像の分割は 9:1 とした．また，画像及び台詞の特徴ベクトルの次元は 768 と設定した．

表 1: 入れ替え推定結果 (台詞入れ替え)

クラス	Precision	Recall	F-1	サンプル数
0	0.45	0.65	0.53	85
1	0.38	0.15	0.22	85
2	0.45	0.52	0.48	85
macro avg	0.43	0.44	0.41	255
base line	0.33	0.33	0.33	
accuracy				0.4392

表 2: 入れ替え推定結果 (画像入れ替え)

クラス	Precision	Recall	F-1	サンプル数
0	0.42	0.36	0.39	85
1	0.47	0.44	0.45	85
2	0.47	0.56	0.51	85
macro avg	0.45	0.45	0.45	255
base line	0.33	0.33	0.33	
accuracy				0.4549

表 3: 入れ替え推定結果 (台詞, 画像入れ替え)

クラス	Precision	Recall	F-1	サンプル数
0	0.38	0.42	0.40	85
1	0.43	0.33	0.37	85
2	0.53	0.60	0.56	85
macro avg	0.45	0.45	0.45	255
base line	0.33	0.33	0.33	
accuracy				0.4510

5.1 実験パターン 1(台詞入れ替え)

4 コマ漫画の 4 コマ目の台詞のみを他の台詞に入れ替えて識別実験をする．クラスとしては入れ替えてないものを 0, 同じ作品の全話中の 4 コマ目の内異なる台詞と入れ替えたものを 1, 異なる作品の 4 コマ目の台詞と入れ替えたものを 2 とした．

5.2 実験パターン 2(画像入れ替え)

4 コマ漫画の 4 コマ目のコマ画像のみを他のコマ画像に入れ替えて識別実験をする．クラスとしては入れ替えてないものを 0, 同じ作品の全話中の 4 コマ目の内異なるコマ画像と入れ替えたものを 1, 異なる作品の 4 コマ目のコマ画像と入れ替えたものを 2 とした．

5.3 実験パターン 3(台詞，画像入れ替え)

4 コマ漫画の 4 コマ目のコマ画像と台詞の両方を他のコマ画像と台詞セットで入れ替えて識別実験をする．クラスとしては入れ替えてないものを 0, 同じ作品の全話中の 4 コマ目の異なるコマ画像と台詞をセットで入れ替えたものを 1, 異なる作品の 4 コマ目のコマ画像と台詞をセットで入れ替えたものを 2 とした．

5.4 結果

表 1, 2, 3 にそれぞれの実験パターンにおいて得られた識別結果を示す．

表 1, 2, 3 から，識別率はベースライン 33% に対して提案手法が約 44% から 46% と上回ったことがわかる．また，本実験において全体的にクラス 1 の F-1 値が低くなる傾向がみられた．これはクラス 1 において，同じ作品という制約の中で異なる台詞，

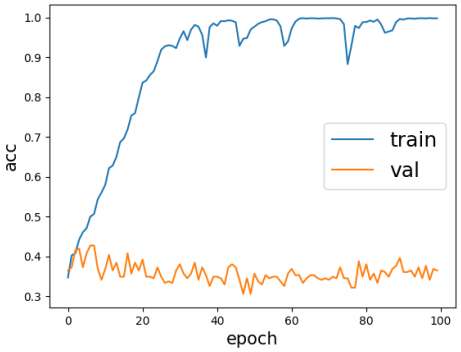

図 2: 識別率の推移 (台詞入れ替え)

画像と入れ替えたために識別境界を決定するのが困難となったために識別率が下がったと考察できる．また, 画像と台詞を両方入れ替えた実験においては, それぞれを個別に抜いた実験の識別率の特徴を合わせたような値が得られていることから，本実験においては同じコマ同士の対応に重点が置かれており，4 コマ漫画のストーリーの繋がりとマルチモーダル性を組み合わせて有効に扱うまでには至らなかったということが考察できる．しかしながら，クラス 2 の F-1 値が個別に抜いたものよりも高くなっていることから，マルチモーダル的な学習の可能性が示されたといえる．

次に，図 2, 3, 4 に訓練および評価識別率の推移を示す．横軸はエポックを，縦軸は識別率を表す．

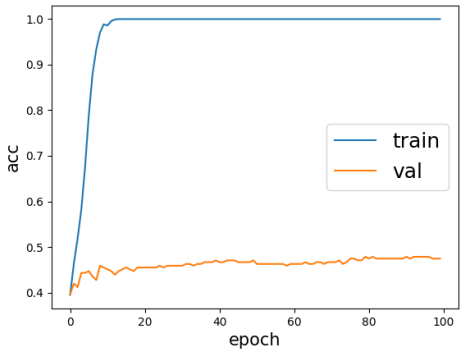

図 3: 識別率の推移 (画像入れ替え)

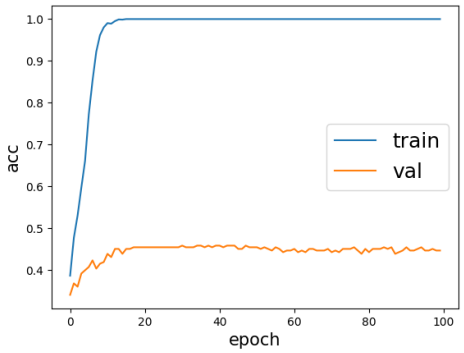

図 4: 識別率の推移 (台詞，画像入れ替え)

図 2, 3, 4 から，訓練識別率の増加に対して評価識別率の増加が抑えられており，20 エポックほどで収束していることがわかる．これは 4 章で述べた「画像と台詞の任意性」とも関係し，訓練によって得られた識別境界が評価画像に十分には適合しないためであると考えられる．また，本実験において，台詞の入れ替え識別に対しては識別率の推移の変動が激しいが，画像の入れ替え識別に対しては比較的安定していることが特徴的であった．この結果は画像の入れ替えよりも台詞の入れ替えの方が任意性が強いということを示しており，ある程度の台詞の違いを許容できるモデルを構築することで識別率の向上が見込めるのではないかと考察できる．

6　まとめと今後の課題

本研究において，漫画をマルチモーダルデータとして扱う研究の取り掛かりとして，4 コマ漫画のコマ画像と台詞の対応識別タスクに取り組んだ．対応識別実験においてはベースラインを上回る識別率が得られた．また，識別率の推移の観察により，漫画における画像と台詞の任意性の差異を確認することができた．本研究では画像と台詞の情報をセットで入れ替えると識別率がその中間程度の値をとったことから，本研究で用いたモデルでは時系列の繋がりまでは見られておらず，時系列情報とマルチモーダルデータをうまく扱えなかったということが分かった．本実験では訓練識別率に対して想定通りに評価識別率が上がらなかったが，グリッドサーチを用いたモデルパラメータの改善やデータの増量および整形などにより識別率の向上が期待できる．

今後の課題としては本実験で識別率の下がっていた同作品間からの置換をしたクラスのより効率的な分離，画像および台詞の任意性を柔軟に考慮したうえで学習するモデルの構築，および分散表現を用いない End to End なモデルを用いた識別実験などが挙げられる．

参考文献

[1] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

[2] Toru Ogawa, Atsushi Otsubo, Rei Narita, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object detection for comics using manga109 annotations. *CoRR*, abs/1803.08670, 2018.

[3] 上野 未貴. 創作者と人工知能: 共作実現に向けた創作過程とメタデータ付与 4 コマ漫画ストーリーデータセット構築. **人工知能学会全国大会論文集**, JSAI2018:4Pin116–4Pin116, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.