深層学習に基づく4 コマ漫画の感情推定とマルチモーダル化への検討

A Method of Emotional Estimation and An Investigation of Multimodalization about Four-Scene Comics Based on Deep Learning

高山 裕成 [1] 岡田 真 [1] 森 直樹 [1]
Yusei Takayama Makoto Okada Naoki Mori

1 はじめに

人工知能の基盤である深層学習を始めとする機械学習技術の大きな発展を受けて, 人工知能を用いた創作物理解が注目されている. しかし, 創作は高次の知的活動であるため, いまだに難しいタスクである. 人の創作物の理解に関する分野の中でもコミック工学など漫画を対象とした研究は, 絵と文章から構成される漫画を対象とするため, 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データを扱う分野である. コミック工学の分野では様々な研究が報告されているが, その多くは画像処理に基づいた研究であり, 自然言語処理による内容理解を目指した研究は少ない. その一因はデータにある. 漫画に含まれるテキストには, 口語表現, 擬音語, 表記揺れといった漫画特有の言語表現を含み, これらの扱いについて考慮する必要がある. そして, 漫画が著作物であることに起因する研究用データの不足も課題となっている.

本稿では人工知能を用いた漫画の内容理解のために, まず自然言語処理を用いた漫画のセリフの感情を推定して, その上で漫画のコマの画像情報を加えたマルチモーダル化について検討する.

2 研究用コミックデータ

4 コマ漫画を対象としたデータセットとしては Manga 109[1] が知られているが, 漫画に登場するキャラクタの感情は明示されていない. そのために人手によるアノテートでラベルを付与する必要があるが, アノテートされたラベルが漫画家の意図とは異なる恐れがある. そこで, 本実験では上野によって作られたデータセット [2] (以下, 4 コマ漫画ストーリーデータセットとする) を用いる. この 4 コマ漫画ストーリーデータセットは同一プロットの下, 幾人かの漫画家によって描き下ろされた 4 コマ漫画で構成されており, 作者によって感情ラベルがアノテートされている. また, 上野は異なる作者によって描かれた 4 コマ漫画を, 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, 萌えタッチと分類している. 図 1 にこのデータセットにおける 2 つのタッチのコマ画像の例を示す. 図 1 より, 同一プロットであっても, 作者の感性によってアノテートされたラベルが異なっていることが分かる.

[1]大阪府立大学, Osaka Prefecture University

図 1: タッチごとのアノテート例

3 要素技術

3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers)[3] は 2018 年に Google が発表した言語モデルの一種である.

その名前の通り, Transformer により構成される双方向のエンコーダであり, 文書分類や質疑応答といった様々な自然言語処理の幅広いタスクにおいて公開時点での最高性能を達成している. 従来のニューラルネットワークを用いた自然言語処理モデルは, 特定のタスクに対して 1 つのモデルを用いてきたが, BERT は転移学習により, 1 つのモデルで, 様々なタスクに対応できる. また, BERT を転移学習したモデルは, 少ないデータを追加学習するのみで動作するため, 1 からモデルを構築しなくて済む. 本稿では日本語 Wikipedia より全 1800 万文を用いて事前学習させたモデル[1] を用いた.

3.2 illustration2vec

illustration2vec[4] は Saito, Matsui らが提案した画像のベクトル化手法であり, Danbooru と Safebooru から 100 万枚のイラストを用いて学習した事前学習済みモデルが公開されている. illustration2vec が扱った問題として, イラストに対する画像認識の難しさがあり, 既存の画像認識モデルのほとんどが ImageNet などの実画像を評価対象にしており, アニメや漫画といったイラストに対して評価をしていなかった. illustration2vec はそれらと比較してイラストのより合理的なベクトル化が期待できる手法である. 本稿では事前学習済みモデルを使って 4096 次元のコマ画像のベクトルを獲得する.

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php

表 1: データ数

	ラベル	ギャグ	少女	少年	青年	萌え
train&valid	喜楽	15	39	15	18	25
	その他	40	26	45	44	35
	喜楽 (拡張後)	1115	2575	940	998	1766
	その他 (拡張後)	2851	1391	3076	3145	2323
test	喜楽	10	38	12	14	22
	その他	56	29	52	51	42

4 Data Augmentation

4 コマ漫画ストーリーデータセットの欠点として, データ数が少ないことがあげられる. そこで, 本実験では日本語 WordNet [5] のシソーラスを用いてテキストデータを拡張する. 分かち書きされたオリジナルのセリフに対して, 日本語 WordNet で類似語を持つ単語について類似語に置き換え, 文を生成することでテキストデータを拡張した. ただし, 文の中に類義語を持つ単語が複数あった場合, 類似語に置き換える単語は同時に 1 つまでとし, 英数字・記号のみで表されている類似語は除外した. 例えば, 5 つの単語からなる文章があり, 各単語が 5 つの類似語を持っている場合, その文からは新しく 25 文が生成されることとなる.

5 実験

本研究では, 各タッチについて感情推定をした. 使用するデータセットには全 7 種類の感情ラベル（ニュートラル, 驚愕, 喜楽, 恐怖, 悲哀, 憤怒, 嫌悪）が含まれているが, データ数と解析の難しさの問題から, 今回は喜楽のみを正例, その他を負例とする 2 クラスに分類した.

訓練用データは各タッチの前半 1 話から 5 話までの拡張されたセリフを用い, 評価用データは後半 6 話から 10 話におけるオリジナルのセリフのみを用いた. 表 1 に各実験で用いたデータ数を示す.

多くのタッチにおいて, 正例は負例に対してデータ数が非常に少ない不均衡データであることから, 損失関数に使うクラス重みとして各タッチについて, 訓練用データの各ラベルのデータ数の逆数を正規化したものを用いた. また, 訓練用データの内 20 % をサンプリングして検証用データとし, 検証用データにおける正例の F1 値が最大となる epoch の結果を評価用モデルとして採用してモデルの性能を測った.

5.1 実験 1：セリフ 1 文の感情推定

実験 1 では 1 つのセリフを入力し, 対応する感情ラベルを出力するような感情推定をした. 識別器としては 3 層 MLP を用いた. 表 2 に MLP で用いたパラメータ, そして表 3 に学習で用いたパラメータを示す. 学習率は Optuna[6] によって最適なパラメータを探索した. JUMAN++ [2] によって分かち書きされたセリフをさらにサブワード化したものを BERT への入力とした.

BERT の事前学習済みモデルの全ての重みを固定した場合 (BERT fixed) と最終層のパラメータだけをチューニングした場合 (BERT fine tuning) の 2 つの手法について感情推定をして, 各タッチ・手法の結果について比較し, fine tuning の有用性を確かめた.

表 2: 実験 1・3 MLP パラメータ

parameters	実験 1	実験 3
(in,hidden,out)	(768,30,2)	(4864,300,2)
activation function	tanh	
dropout rate	0.5	

表 3: 学習パラメータ

parameters	実験 1・2・3
epoch	200
batch size	16
loss function	Cross Entropy Loss
optimizer	Adam

5.2 実験 2：過去のセリフを考慮した感情推定

実験 2 では, 連続したセリフ n 文を入力し, 末尾のセリフに対応する感情ラベルを出力する感情推定をした. 図 2 に $n = 3$ とした時のネットワークの概略図を示す. $\{s_i\}$ はそれぞれ連続する n 文のセリフを BERT の 単語 ID 列に変換したものである. 単語 ID 列はセリフ内の最大単語長に合わせて, 後方パディングをしている. また, 4 コマ漫画ストーリーデータセットは 1 話につき 2 つの 4 コマを含んでいるが, 各 4 コマは時系列的に繋がっていないことを踏まえて, s_i のセリフの組み合わせとしては同一の 4 コマに属し, かつ連続しているものを扱う. 各 4 コマの序盤に現れるセリフには参照できる過去のセリフが無いため, 便宜上のセリフ “[pad]” を置くことで対処した. また, BERT への入力時における 1 文目, 2 文目を表す ID は奇数番目の入力では 0, 偶数番目の入力では 1 となるようにした. そして, 末尾のセリフ以外はオリジナルのセリフのみから抽出し, すべての組み合わせを入力列とすることで, 実験 1 のデータ数と合わせた.

単語 ID 列長を w とすると入力次元は $(batch \times n \times w)$ となる. このままでは BERT の入力次元に対応していないので, まず $(n \times batch \times w)$ へと軸を入れ替え, これを 1 次元目について各ベクトルに分解し, これら n 個の 次元数 $(batch \times w)$ のベクトルをそれぞれ BERT への入力とし, BERT の出力から [CLS] トークンに相当するベクトルのみをそれぞれ抜き取り, 先と逆の手順を踏むことで次元数 $(batch \times n \times 768)$ のテンソルを得る. これを識別器として用いる Bi-LSTM ベースの Self-Attention への入力とすることで末尾のセリフの感情を推定するように学習を行う. 本稿では n を 2 から 5 に変えて実験した. また, BERT は最終層のみをチューニングした.

表 3 に学習で用いたパラメータ, 表 4 に識別器のパラメータを示す.

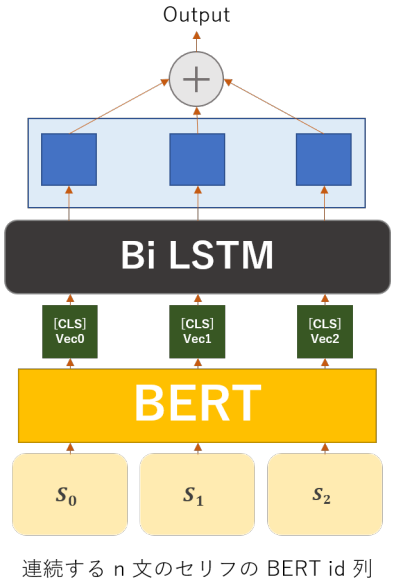

図 2: 提案ネットワーク $(n = 3)$

[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

表 4: 実験 2 識別器パラメータ

parameters	value
lstm_in	768
lstm_hidden	128
self_atten_in	128×2
atten_num_layers	3

5.3 実験 3：マルチモーダルな感情推定の検討

実験 3 では, BERT から得た 768 次元のセリフベクトルと, 入力したセリフが含まれているコマ全体の画像を illustration2vec に入力して得た 4096 次元のコマベクトルを concat した 4864 次元のベクトルを入力ベクトルとして, 3 層 MLP に入力することでマルチモーダルな感情推定をした. また実験 1 と同様にして, BERT の事前学習済みモデルの全ての重みを固定した場合 (Multi fixed) と最終層のパラメータだけをチューニングした場合 (Multi fine tuning) の 2 つの手法について感情推定をして, 各タッチ・手法の結果について比較した.

6 実験結果

実験 1, 2, 3 において, すべての出力が負例と推定された場合の評価指標の値をベースラインとして設定した. 以下, 表における P-Recall, P-F1 値は正例の再現率, F1 値を表し, Acc は全体の精度を表す. また, Recall が 0 の時の F1 値は 0 とした. 5 タッチ総合の結果は各タッチにおける混合行列の和から算出してある.

6.1 実験 1：結果

表 5 に実験 1 の結果を示す. 表 5 より, Accuracy に関して BERT fine tuning のみがベースラインを超え, すべての評価指標において BERT fixed より BERT fine tuning の方が上だったことから, fine tuning の有用性を確かめることができた.

6.2 実験 2：結果

表 6 に実験 2 の結果を示す. 表 6 より, 実験 1 と比べ, 全体的に評価指標は下回ったことから, 過去のセリフを考慮することの優位性を示すことはできなかった. その理由としてはモデルの設計上, BERT から得た分散表現が文同士の繋がりを学習できておらず, 逆に BERT の強みを潰してしまっている可能性があることが挙げられる. しかし, n が大きくなるに従って P-Recall, P-F1 値は増加傾向にあり, P-Recall に関しては n が 3 以上の時, 実験 1 の結果を上回っていることから過去のセリフをより多く考慮することの有用性はあると推測できる.

6.3 実験 3：結果

表 7 に実験 3 の結果を示す. 表 7 より, まず実験 1 と同様にすべての評価指標において Multi fixed より Multi fine tuning の方が上だったことから, BERT の fine tuning の有用性を確かめることができた. また, 実験 1 と比べて Accuracy は下回ったものの, P-Recall, P-F1 値は増加し, 特に少女漫画タッチと少年漫画タッチにおいては実験 1 と比較すると最も高い P-Recall と P-F1 値が得られており, マルチモーダルな感情推定の意味があると推測できる.

7 考察

まず, 各実験の結果についての考察を述べる. 過去のセリフを考慮することで 実験 2 において, 全体的に評価指標が下回ったのは, ネットワークの柔軟性が上がり, より過学習が進んだことで汎化性能が下がったことが原因だと考えられる. Optuna で探索する学習率の範囲は共通しており, 学習曲線を見ても実験 2 の方がより早く収束している場合が多かったことから, 柔軟性に合わせて探索範囲を変えたり, 学習率スケジューラーを設定したり, 適宜 Dropout 層を追加するなどをして過学習を抑制する必要があると考えられる. 実験 3 においても同様のことが言えるが, コマの画像ベクトルがセリフの分散表現に対して次元数がかなり多いことを踏まえると, 全結合層などを追加し, 次元数を合わせる必要もあると考えられる. ただし, illustration2vec のモデルがこの実験に適しているかの妥当性は確認していないため, これは今後の課題とする.

次に, データ及びモデルの問題点について考察する. 本研究では Data Augmentation の手法における文法的齟齬の問題については取り組んでおらず, この問題がどれだけ寄与しているかについての調査は今後の課題である. 具体的な解決策としては, 拡張されたセリフの分散表現とそれぞれに対応するオリ

ジナルのセリフの分散表現とのコサイン類似度から閾値未満のデータを除外したり, 拡張する品詞を限定するといったことが考えられる. また, 本稿で用いた BERT の事前学習済みモデルでは, 'A' や 'B' 等のアルファベットは未知語として扱われている. 4 コマ漫画ストーリーデータセットにおいてはこれらの記号が主役であるキャラクタ名に充てられていることから, 主語の情報が抜け落ちてしまい, 会話の流れを上手く学習できていない可能性があることから, 発話者情報を加味したり, セリフ内のキャラクタ名を特殊なトークンに置き換えるなどの工夫が必要であると言える. そして, そもそものデータ数が極めて少ない問題の解決も重要な課題である.

8 まとめと今後の課題

本研究では BERT で得たセリフの分散表現からセリフの感情推定をした. 実験結果から過去のセリフを考慮することの優位性を示すには多くの課題があることが分かった. また, コマの画像のベクトルも考慮したマルチモーダルな感情推定から, 1 つのセリフを入力とする実験においては優位性を確認した. 時系列を考慮したセリフのマルチモーダルな感情推定は実験 2 において入力したそれぞれのセリフが含まれているコマ全体の画像のベクトルを BERT からの出力にそれぞれ concat すれば可能であり, これは今後の課題とする.

そして, 4 コマ漫画ストーリーデータセットのみでの訓練・テストでは精度の向上が困難であることから, 何かしらの日本語の極性評価データセットを用いて BERT の事前学習モデルを fine tuning し, 4 コマ漫画の感情推定問題に転移させることが考えられる. 場合によっては感情ラベルの貼り直し, または人手によるラベル付けが必要になる. また Manga 109 のセリフのデータを用いることで半教師あり学習も可能であり, マルチモーダルな感情推定手法についても更なる工夫が必要である.

なお, 本研究は一部，日本学術振興会科学研究補助金基盤研究 (B) (課題番号 19H04184)，基盤研究 (C) (課題番号 20K11958) の補助を得て行われたものである.

表 5: 実験 1 結果 (評価用データ)

	ギャグ			少女漫画			少年漫画			青年漫画			萌え			5 タッチ総合		
	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1
BERT fixed	0.712	0.200	0.174	0.567	0.632	0.623	0.766	0.083	0.118	0.692	0.643	0.474	0.594	0.591	0.500	0.666	0.510	0.473
BERT fine tuning	0.818	0.200	0.250	0.612	0.711	0.675	0.766	0.083	0.118	0.862	0.500	0.609	0.609	0.591	0.510	0.733	0.521	0.535
ベースライン	0.848	0	0	0.432	0	0	0.812	0	0	0.784	0	0	0.656	0	0	0.705	0	0

表 6: 実験 2 結果 (評価用データ)

n	ギャグ			少女漫画			少年漫画			青年漫画			萌え系			5 タッチ総合		
	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1
2	0.712	0.200	0.174	0.493	0.684	0.605	0.781	0.083	0.125	0.815	0.286	0.400	0.531	0.364	0.348	0.666	0.427	0.429
3	0.621	0.400	0.242	0.507	0.763	0.637	0.750	0.167	0.200	0.785	0.357	0.417	0.516	0.591	0.456	0.635	0.552	0.471
4	0.636	0.300	0.200	0.537	0.816	0.667	0.766	0.167	0.211	0.800	0.357	0.435	0.500	0.591	0.448	0.647	0.563	0.484
5	0.652	0.500	0.303	0.507	0.895	0.673	0.781	0.333	0.364	0.754	0.500	0.467	0.594	0.591	0.500	0.656	0.656	0.529
ベースライン	0.848	0	0	0.432	0	0	0.812	0	0	0.784	0	0	0.656	0	0	0.705	0	0

表 7: 実験 3 結果 (評価用データ)

	ギャグ			少女漫画			少年漫画			青年漫画			萌え			5 タッチ総合		
	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1	Acc	P-Recall	P-F1
Multi fixed	0.758	0.000	0.000	0.627	0.553	0.627	0.391	0.750	0.316	0.554	0.500	0.326	0.563	0.273	0.300	0.580	0.448	0.386
Multi fine tuning	0.773	0.200	0.211	0.687	0.763	0.734	0.703	0.417	0.345	0.769	0.643	0.545	0.641	0.500	0.489	0.715	0.583	0.546
ベースライン	0.848	0	0	0.432	0	0	0.812	0	0	0.784	0	0	0.656	0	0	0.705	0	0

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] 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**, 2018.

[3] Chang M.-W. Lee K. Devlin, J. and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[4] Yusuke Matsui Masaki Saito. Illustration2vec: A semantic vector representation of illustrations. 2015.

[5] Richard Fothergill Francis Bond, Timothy Baldwin and Kiyotaka Uchimoto. Japanese semcor: A sense-tagged corpus of japanese in the 6th international conference of the global wordnet association (gwc-2012), matsue. 2012.

[6] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.