

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

1 はじめに

近年, 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が大きく発展してきている. また, その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が, 創作は高次の知的活動であるため, 未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な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, これらを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準備として自然言語処理を用いた漫画のセリフの感情を推定を行い, 更には漫画のコマの画像情報を加味したマルチモーダル化への検討を行う.

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 が知られているが, 漫画に登場するキャラクターの感情は明示されていない. そのため人手によるアノテーションでラベルを付与する必要があるが, アノテーション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 [1] (以下, 4コマストーリーデータセットとする) を用いる. この4コマ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ションされている. 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



図1: タッチごとのアノテーション例

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ションされたラベルが異なっていることが分かる.

3 要素技術

3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers)[2] は2018年にGoogleが発表した言語モデル一種である.

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ーであり, 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つのモデルをチューニングすることで, 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ので, 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1,800万文を用いて事前学習させたモデル¹を用いた.

¹<http://nlp.ist.i.kyoto-u.ac.jp/index.php>

3.2 illustration2vec

illustration2vec[3] は, Masaki Saito と Yusuke Matsui らが提案した画像のベクトル化手法で, 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 で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行ってなかったため, イラストのベクトル化手法に関してより良い期待が持てる手法である. 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.

4 Data Augmentation

4 コマ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 [4] のシソーラスを用いてデータを拡張する手法を用いる. 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を類似語に置き換えた. 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

本実験では, 各タッチについて感情推定を行った. 本実験で使用するデータセットは全 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) を持っているが, 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用い

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

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 F 値が最大となる epoch のものを評価用モデルとして採用してモデルの性能を測るものとする.

5.1 実験 1: 時系列を考慮しないセリフの感情推定

この実験では 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行う. JUMAN++² によって分かち書きされたセリフを BERT への入力とし, 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 によって最適なパラメータを探索した.

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 last layer) の 2 つの手法について感情推定を行い, 各タッチ・手法の結果について比較し, fine tuning の有用性を確かめることを目的とした.

表 2: 実験 1 MLP パラメータ

	MLP
(in,hidden,out)	(768,30,2)
activation function	tanh
dropout rate	0.5

表 3: 学習パラメータ

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

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

5.2 実験 2：時系列を考慮したセリフの感情推定

この実験では、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ような感情推定を行う。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、この 4 コマ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。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 Bi LSTM 層、Self-Attention 層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。

BERT は最終層のみをチューニングする。学習で用いたパラメータは表 3、識別器のパラメータは図 4 に示す。

本稿では n が 2 から 5 の場合について実験を行った。

表 4: 実験 2 識別器パラメータ

parameter	value
lstm_in	768
lstm_hidden	128
self_attn_in	128×2
atten_num_layers	3

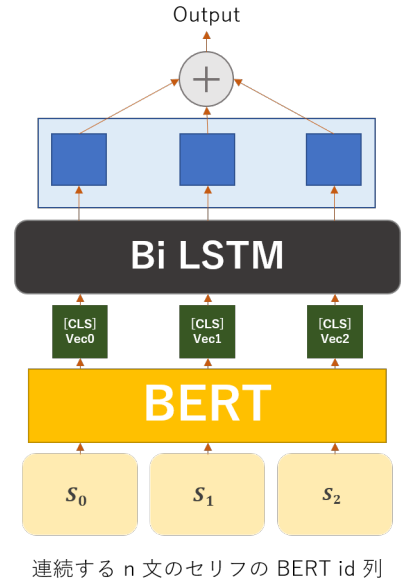


図 2: seq net ($n = 3$)

5.3 実験 3：単文のマルチモーダル感情推定

実験 1 において、BERT から得た 768 次元のベクトルと、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、得た 4096 次元のベクトルを concat した 4864 次元のベクトルを、同様に 3 層 MLP に入力することでマルチモーダルな感情推定を行った。データセット整備の都合上、少年漫画タッチ、青年漫画タッチ、萌えタッチについてのみ実験を行った。

6 実験結果

実験 1, 2, 3 において、すべての予測値が負例である場合の各評価指標の値をベースラインとして設定した。以下、図における Recall, F1 値は正例のものを表し、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 は各タッチについて、評価用データの結果をまとめたものである。表 5 より、Accuracy に関しては差異はないが共にベースラインを超え、Recall, F1 値共に BERT fixed より BERT last layer の方が上回ったことから、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 は各タッチについて、評価用データの結果をまとめたものである。また表 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、時系列を考慮することの有用性を示すことは出来なかったが、 n が大きくなるに従って F1 値は増加傾向にあることから時系列を考慮する意味はあると推測できる。

6.3 実験 3：結果

表 7 は 3 タッチについて、評価用データの結果をまとめたものである。比較として、実験 1 における 3 タッチの結果も載せている。表 7 より、実験 1 と比べ、Recall 以外の評価指標は下回ったが、少年漫画タッチにおいては今回の実験を通して最も高い Recall と F1 値が出ており、マルチモーダルな感情推定を行う意味はあると推測できる。

7 考察

まず、各実験の結果についての考察を述べる。実験 2 において、時系列を考慮することで実験 1 と比べ、全体的に評価指標は下回ったのは、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、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推測できるが、未だ 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。本稿では 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は今後の課題とする。具体的な解決策としては、拡張されたセリフとそれぞれに対応するオリジナルのセリフ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挙げられる。また、本稿で用いた BERT の事前学習済みモ

デルでは、‘!’ や ‘?’ は未知語となっており、漫画においてはこれらの記号がニュートラル以外の感情を表す表現であるという経験則や、実際にデータの中にこれらの記号のみで構成され、ニュートラル以外の感情ラベルが付与されているセリフがあることから、特殊なトークンに置き換えて学習させたり、またはこれらの記号が辞書に含まれている他の事前学習済みモデルを用いて精度を比較する必要があると考えられる。そして、そもそものデータ数が極めて少ないことが問題として挙げられる。

8 まとめと今後の課題

本稿では 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。時系列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も行ったが、1 つのセリフを入力とする実験では優位性は見られなかった。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 すれば可能であるから、直近の課題とする。

今後の実験の方針として、4 コマストーリーデータセットのみでの訓練・テストでは精度に期待できないことから、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、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 manga109 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ると考えられる。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と考えられる。

参考文献

- [1] 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.
- [2] Chang M.-W. Lee K. Devlin, J. and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.
- [3] Yusuke Matsui Masaki Saito. Illustration2vec: A semantic vector representation of illustrations. 2015.
- [4] 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.

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			萌え系			5 タッチ総合		
	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
BERT fixed	0.757	0.200	0.200	0.552	0.684	0.634	0.796	0.000	0.000	0.784	0.357	0.416	0.718	0.454	0.526	0.720	0.447	0.485
BERT last layer	0.818	0.200	0.250	0.641	0.710	0.692	0.781	0.000	0.000	0.800	0.500	0.518	0.578	0.500	0.448	0.723	0.489	0.510
ベースライン	0.848	0.000	0.000	0.432	0.000	0.000	0.812	0.000	0.000	0.784	0.000	0.000	0.656	0.000	0.000	0.705	0.000	0.000

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	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
2	0.803	0.200	0.235	0.492	0.526	0.540	0.765	0.083	0.117	0.738	0.285	0.320	0.578	0.363	0.372	0.674	0.364	0.397
3	0.772	0.000	0.000	0.626	0.526	0.615	0.765	0.000	0.000	0.800	0.285	0.380	0.625	0.409	0.428	0.717	0.343	0.417
4	0.772	0.000	0.000	0.582	0.578	0.611	0.781	0.083	0.125	0.769	0.571	0.516	0.546	0.681	0.508	0.687	0.479	0.474
5	0.742	0.500	0.370	0.641	0.552	0.636	0.781	0.000	0.000	0.769	0.428	0.444	0.625	0.500	0.478	0.711	0.447	0.477
ベースライン	0.848	0.000	0.000	0.432	0.000	0.000	0.812	0.000	0.000	0.784	0.000	0.000	0.656	0.000	0.000	0.705	0.000	0.000

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			萌え系			3 タッチ平均		
	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
実験 1 fixed	-	-	-	-	-	-	0.796	0.000	0.000	0.784	0.357	0.416	0.718	0.454	0.526	0.766	0.312	0.400
実験 1 last layer	-	-	-	-	-	-	0.781	0.000	0.000	0.800	0.500	0.518	0.578	0.500	0.448	0.720	0.375	0.400
実験 3	-	-	-	-	-	-	0.562	0.500	0.300	0.784	0.500	0.500	0.562	0.363	0.363	0.637	0.437	0.375
ベースライン	-	-	-	-	-	-	0.812	0.000	0.000	0.784	0.000	0.000	0.656	0.000	0.000	0.751	0.000	0.000