卒 業 研 究 報 告 書

題　　目

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

研 究 グ ル ー プ　　第 1 研究グループ

指 導 教 員　　森 直樹 教授

令 和　2　年（　　2020　年）度 卒 業

（No.　　1171201102　　）　　高山 裕成

大阪府立大学工学域電気電子系学類情報工学課程

目 次

図 目 次

表 目 次

1　はじめに

近年, 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, 創作物理解や作品の自動生成といった試みは工学的に興味深く意義が大きい反面，そもそも人の創作物理解は高次の知的活動であり，どういったタスクであれば人工知能が創作物を理解したと言えるのかを定義することさえ現状では難しい. 人の創作物の理解に関する分野の中でもコミック工学 [4] 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. その大きな原因の一つとして, 漫画が著作物であることに起因するデータ不足が挙げられる. 漫画という媒体の情報を十分に活用するにはマルチモーダルな解析が最も良いと考えられ, その中でも大きな意義があると考えられるタスクはセリフの感情推定である. その理由としては 2 つある. 人工知能を用いた対話型システムの精度向上のためには自然な表情や発言, 振る舞いから人の潜在的な感情を推定できるようになる必要があること. そして, 創作者に対して作品の展開を汲んだ適切なセリフの自動生成といった制作時間の削減や質の向上に関わる創作支援に繋がるからである.

したがって, 本研究では人工知能を用いた漫画の内容理解のために, 漫画におけるキャラクタのセリフのマルチモーダルな感情推定を目的とする. 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. また, データ不足を解決するためにシソーラスを用いたデータ拡張をする.

以下に本論文の構成を示す. 2 章ではコミック工学に関連するデータセットについて説明し, 3 章では本研究に関連する要素技術と従来研究について概説する．4 章では漫画のセリフのマルチモーダルな感情推定のための提案手法とシソーラスを用いたデータ拡張について述べる. そして, 5 章において実験とその考察を示す. 最後に 6 章で本研究の成果をまとめた上で, 今後の課題について述べる.

2　コミック工学に関連するデータセット

本章では, コミック工学に関連するデータセットについて説明する.

2.1　Manga109

Manga109 [5] は, 漫画の研究のために相澤らにより作られたデータセットである. このデータセットは日本のプロの漫画家によって描かれた 109 冊の漫画で構成されている. これらは 1970 年代から 2010 年代に公開された漫画であり, 対象読者層やジャンルも幅広く網羅されている. Manga109 には 109 冊の漫画の画像データ, 登場人物の名前, 画像内における登場人物の顔, 全身, コマ, テキストの座標などのアノテーションデータが含まれている. しかし, Manga109 にはセリフの発話者に関する情報や本研究の趣旨であるセリフの感情に関する情報は付与されていない.

2.2　4 コマ漫画ストーリーデータセット

4 コマ漫画ストーリーデータセット [6] は, 上野によって作られたコミック工学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のデータセットである. このデータセットは画像データのレイヤー分けや作者によるセリフの感情アノテーションなどの特徴を持つ.

Manga109 といった市販コミックによって構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のデータは本データセットのために幾人かの漫画家によって描き下ろされている. 市販のコミックをデータとした場合, 著作権などの問題に加えて, 計算機上で扱うためのデータが少なく, コミックの意味理解を目的とした研究には適用が難しいという問題がある. 例えば Manga109 のデータでは, コミックに登場するキャラクターの感情は明示されていないため, 読者によるアノテートによって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可能性を否定できない.

また, マルチモーダルでストーリーの解析をする際にオリジナリティの観点から同一プロットを複数の漫画家が描くことは稀有なため, そういったデータの収集に基づく研究は困難である. 4 コマ漫画ストーリーデータセットはそういった問題点を解決するために作られたデータセットである.

上野は 4 コマ漫画の構造を,

- 一般：標準的な起承転結をもつ
- 繰り返し：1, 2 コマ間の類似が 3, 4 コマ間でも起きる
- 出オチ：1 コマ目におかしな絵が描かれてオチがある
- タイトルオチ：最後にタイトルを見返してオチがわかる
- 再帰：4 コマ目から 1 コマ目に戻り話として成立する
- 参照：1 つ以上前の話の続きの話となる
- 連続した 4 コマを 2 話並べて 8 コマで話となる

と 7 種類に定義し, これに従ってデータセットを作成している. 現在は, 同一のストーリーを 4 コマ目がオチとなる「一般」と「出オチ」の 2 種類の構造から描いたデータがデータセットに納められている.

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

という 5 種類に分類した. また, このデータセットには各タッチにつき 10 話ずつ納められている. 図 2.1 に, 4 コマ漫画ストーリーデータセットのデータ例を示す.

図 2.1: 4 コマ漫画ストーリーデータセットのデータ例

ギャグタッチ：© 作画：浦田カズヒロ

少女漫画タッチ：© 作画：高科りさ

少年漫画タッチ：© 作画：鈴木市規

青年漫画タッチ：© 作画：湯沢としひこ

萌えタッチ：© 作画：棟田ウメコ

シナリオ：(株) スポマ播村早紀 / 大阪工業大学 上野未貴

3　要素技術と 4 コマ漫画に関する従来研究

本章では，本研究に関連する要素技術について説明する．また, 4 コマ漫画に関する従来研究について概説する．

3.1　自然言語処理に関する要素技術

自然言語の単語や文を計算機上で表現するための分散表現獲得手法について説明する．

3.1.1　形態素解析

形態素とは日本語などの自然言語において意味を持つ最小の単位のことであり, 文を形態素に分割し, 各形態素の品詞などを判定する技術を形態素解析という．英語の文では, 単語と単語の区切りがほとんどの箇所で明示的に示される．このため、形態素への分割処理は簡単なルールに基づいて行われる場合が多い．一方で, 日本語の文は単語間の区切りが英語ほど明確でないため, 形態素への分割は困難かつ重要である．

形態素解析器としては, MeCab [7] や Juman++ [8] などが存在する．

3.1.2　局所表現, 分散表現

自然言語の単語を計算機上で表現する手法として, 最もシンプルなものが局所表現である．単語の代表的な局所表現の 1 つに One-hot 表現がある．One-hot 表現は単語をベクトルの各次元に 1 対 1 対応させる表現方法である．非常に単純な手法であり, 実装が容易であるという利点がある．一方で, One-hot 表現では語彙数とベクトルの次元数が等しくなるため, 語彙数の増大とともにベクトルの次元数も増大し, ベクトル空間がスパースになってしまう問題がある．また, 各単語がベクトル空間上で等距離に配置されてしまうため, 単語間の意味的な関係性については定義できないことも大きな問題である．

局所表現の問題点を解決するために考案された手法が分散表現である．分散表現は各概念をベクトルの単一次元ではなく複数次元の実数で表す．単語の分散表現は，類似した文脈で使用される単語は類似した意味をもつ，という分布仮説を基盤としている．単語を実数値密ベクトルで表現することにより, 単語間の意味的な関係性をベクトル空間上での類似度として定義できるという大きな利点がある．

3.1.3　Word2Vec, Doc2Vec

Word2Vec [9] は単語の分散表現を獲得する手法の 1 つである．この手法は, 同じ文脈で出現する単語は類似した意味を持つと予想されることに基づいており, 写像されたベクトルは, One-hot 表現のような局所表現と異なり, 単語間の意味を考慮した類似度測定や,「王様」−「男」+「女」=「女王」のような単語間の意味における演算などができるようになる．

Word2Vec では, 自己から周りの単語あるいは周りの単語から自己を予測することにより分散表現を獲得する．前者の手法を Skip-gram といい, 後者の手法を Countinuous Bag-of-Words (CBOW) という．

Doc2Vec [10] は Word2Vec をベースとした, 文書をベクトル空間上に写像して分散表現を得る自然言語処理の手法である．Paragraph ID は各文書と紐づいており, 単語の学習時に一緒にこの Paragraph ID を学習することで文書の分散表現を獲得する．このベクトルを用いると文書間の類似度の算出や文書間での加減算が可能になる．

CBOW を拡張したモデルを Distributed Memory モデルといい, Skip-gram を拡張したモデルを Distributed Bag-of-Words という．

3.1.4　Attention

機械翻訳のタスクに対して考案された, LSTM [11] に代表される Recurrent Neural Network (RNN) を用いる Encoder-Decoder モデルは可変長の文を固定長のベクトルにエンコードするため, 長い入力文になるほど隠れ層のノード数が不足し, 学習が難しくなる問題がある. そこで Bahdanau らにより提案されたのが Encoder 側で入力文の各単語の荷重を決定してエンコードすべき場所を制御する Attention 機構 [1] である. Luong らはこの Bahdanau らによるモデルを単純化したモデルとして, 入力されたすべての単語を使用する Global Attentional Model [12] を提案した. 図 3.1 にその概略図を示す.

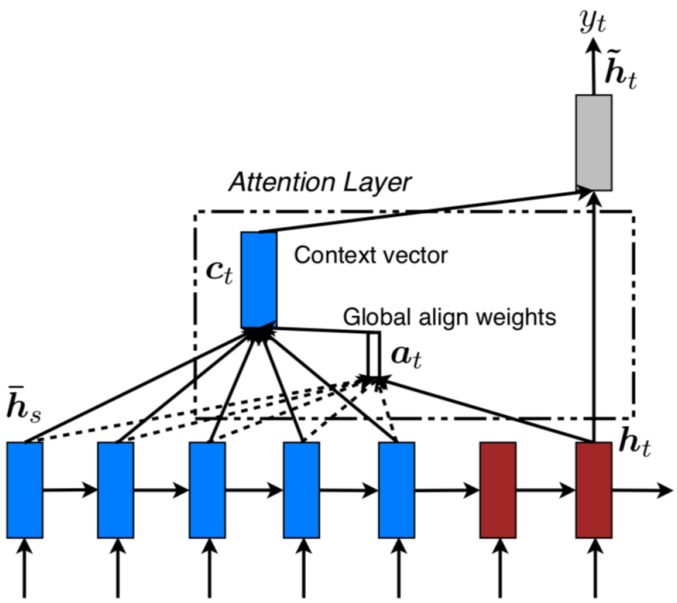

図 3.1: Attention 機構の概略図：文献 [1] の図より引用

3.1.5　Transformer

Transformer [2] は 他モデルで頻繁に用いられてきた RNN を用いずに Attention 機構のみを基本構造とする Encoder-Decoder モデルである．図 3.2 にその概略図を示す．Transformer のエンコーダおよびデコーダはそれぞれ Self-Attention を基本構造とする．Self-Attention とは, Attention 機構の特別な場合である．Attention 機構は Query と Key-Value へのマッピングとして表現することが可能である．通常 Query はデコーダからのターゲットを, Key-Value はエンコーダからのソースを表す．しかし Self-Attention は下層のすべての位置を参照することができシーケンスの依存関係を獲得できる．

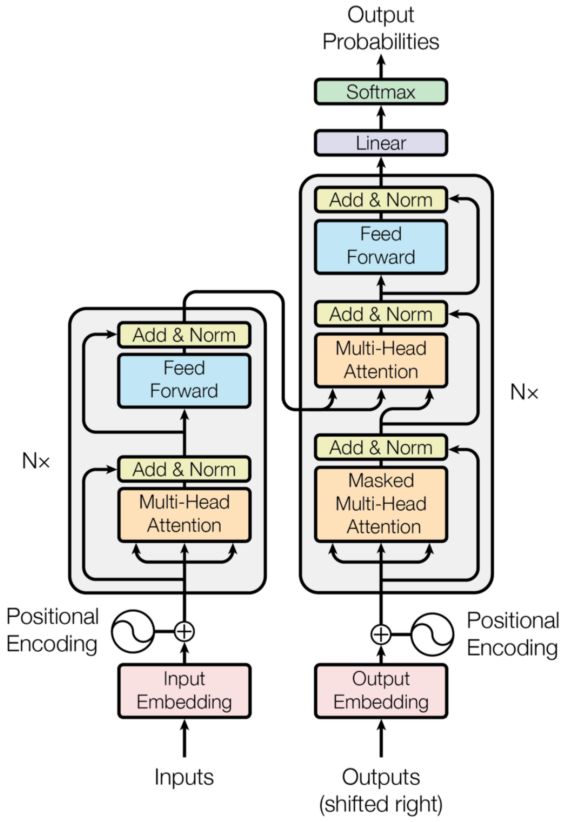

図 3.2: Transformer の概略図：文献 [2] の図より引用

3.1.6　BERT

Bidirectional Encoder Representations from Transformers (BERT) [3] は, 2018 年に Google が発表した言語モデルであり, 複数の双方向 Transformer に基づく汎用言語モデルである. これまでの言語モデルは特定の学習タスクに対して 1 つのモデルを用いてきた. しかし BERT は大規模コーパスに対して事前学習を施して, 各タスクに対して fine-tuning をすることで, さまざまなタスクに柔軟に対応することができる. さらに, 以前はモデルごとに語彙を 1 から学習させるため, 非常に多くの時間とコストがかかっていたが, BERT ではオープンソースで公開されている文脈を既に学習させた Pre-Training BERT モデルを使用することで短時間で学習ができる.

BERT の 事前学習では, 周囲の単語からある単語を予測する Masked Language Model (MLM) と 2 つ目の文章が 1 つ目の文章の次の文章であるかを予測する Next Sentence Prediction (NSP) によりモデルを学習する.

図 3.3 に BERT の事前学習と fine-tuning の概略を示す.

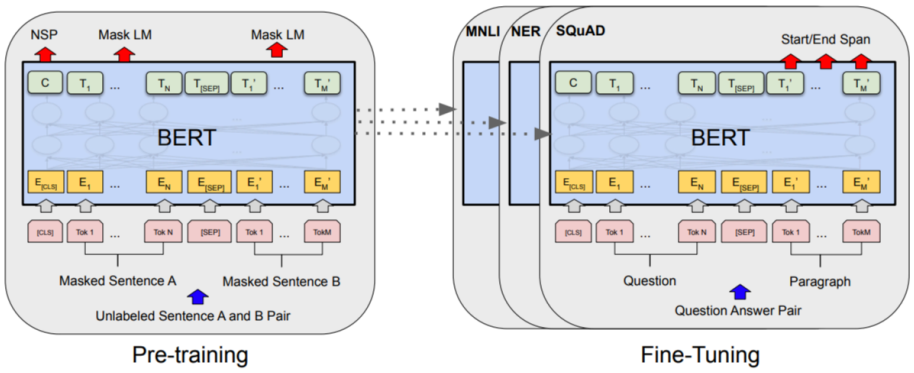

図 3.3: BERT の事前学習と fine-tuning の概略図 : 文献 [3] の図より引用

3.2　画像処理に関する要素技術

画像処理に関する要素技術について説明する.

3.2.1　VGG

VGG [13] とは, ImageNet [14] と呼ばれる大規模画像データセットで学習済みの畳み込みニューラルネットワーク (VGGNet) であり, 13 層の畳み込み層と 3 層の全結合層の合計 16 層からなる. 構成する層の数に応じて, VGG11 や VGG16 などと呼ばれることが多い. VGGNet は Oxford 大学に提案された手法であり, 2014 年の画像認識大会で非常に好成績を収めたことからその後のモデルアーキテクチャに広く取り入れられている. あらかじめ学習が済んでいるモデルを転移学習することで事前学習なしでも深いネットワークを学習できる.

3.2.2　illustration2vec

illustration2vec [15] は齋藤, 松井らが提案した VGG をベースとした画像のベクトル化手法であり, 画像リンク集サイトである 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, VGG などの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 はそれらの手法と比較して, より合理的なイラストのベクトル化が期待できる手法である. また, Danbooru と Safebooru でよく使われているタグを正解ラベルとして学習しているため, 簡単にイラストの特徴を検出でき, 大量の画像に対して類似画像を検索出来たり, 画像の意味における画像変換や応用例としてタグの特徴を満たす画像の生成などが可能となっている.

illustration2vec では学習データを下記の要領で作成している.

1. Danbooru と Safebooru から画像とメタデータを収集する.
2. メタデータを 4 つのカテゴリに分類する.
   - general : 一般的な属性 (例 : “smile”, “short hair”)
   - copyright : 著作権名
   - character : キャラクタ名
   - rating : X レーティング (“safe”, “questionable”, “explicit”)
3. general, copyright, character から最も使われている 512 個のタグをそれぞれ抽出する.
4. 3 で抽出したタグと rating を連結させた 1539 個のタグをラベルとする.

3.3　4 コマ漫画に関する従来研究

4 コマ漫画に関する研究としては，4 コマにおける画像特徴が与える感情識別に関する研究 [16] やストーリー理解過程の解析研究 [17]，4 コマ漫画ではないが既存漫画のデータを利用した 2 コマ漫画の生成に関する研究 [18] が報告されている. また 4 コマ漫画の自動生成に関する研究 [19] や遺伝的アルゴリズムに基づく感性解析に 4 コマ漫画を用いた研究 [20] もなされている.

また，ストーリーに関しては 4 コマ漫画の内容に踏み込んだ研究として，コマの順序識別に関する研究 [21],[22] が報告されている. しかしながら手法，データセットともにまだ十分とは言えず，今後の発展が期待されている分野である.

4　提案手法

本章では，本研究の提案手法について説明する．

4.1　漫画のセリフのマルチモーダルな感情推定手法

本研究では, 4 コマ漫画ストーリーデータセットを用いて, 各セリフにアノテートされた感情ラベルを推定するタスクを解き, その精度を確認する．

図 4.1 にマルチモーダルな推定手法として, 提案手法の概要を示す．図 4.1 のように, ある 4 コマ漫画の中の 1 コマがあって, このコマに含まれている “キャー！” というセリフのラベルを推定することを例に取って説明する．

自然言語処理のみを用いた推定を行う場合は, Text Embedding 層の入力として適した形式にこのセリフを整形し, 出力として得たセリフの分散表現 (以下, “セリフベクトル” と呼ぶ) を識別器への入力とすることで対応する感情ラベルを推定する．そして, マルチモーダルな推定を行う場合は, このセリフが含まれているコマ画像全体を Image Embedding 層への入力とし, 得られたコマ画像の分散表現 (以下, “コマベクトル” と呼ぶ) をセリフベクトルと結合させたベクトルを識別器への入力とすることで対応する感情ラベルを推定する．

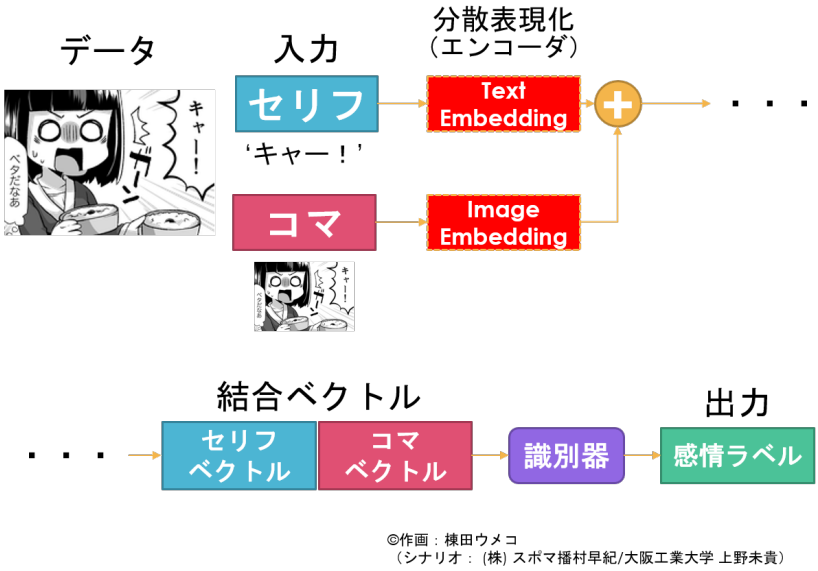

図 4.1: 提案手法の概要

4.2　シソーラスを用いたデータ拡張

4 コマ漫画ストーリーデータセットの問題点として, データ数が少ないことがあげられる. そこで, 本研究では日本語 WordNet [23] のシソーラスを用いてテキストデータを拡張する.

図 4.2 にシソーラスを用いたデータ拡張の概要を示す. 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する. 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する. 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.

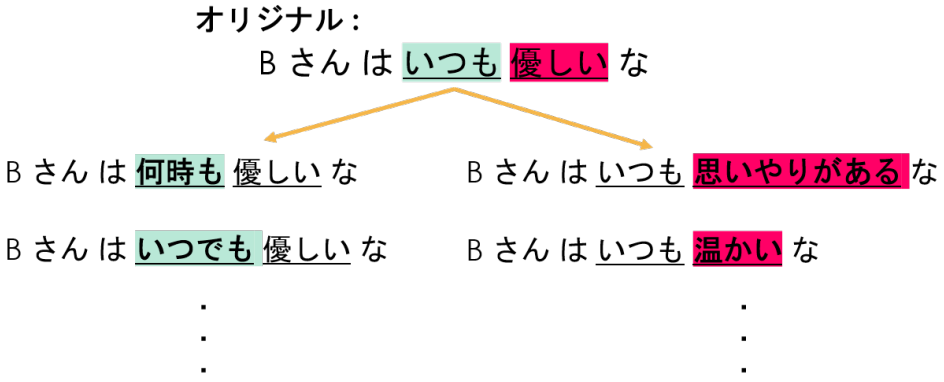

図 4.2: シソーラスを用いたデータ拡張の概要

5　実験と考察

本章では，実験と考察について説明する．本研究では, 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て漫画のキャラクタのセリフの感情を推定する．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な推定手法について実験する．そして, 実験結果からセリフの感情推定とマルチモーダル化の精度への影響について考察する．

5.1　実験に使用したデータ

実験には上野の 4 コマ漫画ストーリーデータセットを用いた．このデータセットには 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）と, アノテーション不備によるラベル不明 (以下, “UNK” とする) の全 8 種類が含まれている．今回は “UNK” のデータを除いた全 7 種類の感情ラベルが付いたデータのみを用いた．表 5.1 に各タッチのオリジナルデータに対する感情ラベルごとのデータ数を示す．

5.2　実験設定

データ数と解析の難しさの問題から, 本研究では 7 種類ある感情ラベルのうち, 喜楽のみを正例クラス, その他の感情ラベルをすべて負例クラスとした 2 クラスのセリフの感情推定を行った．

訓練用データは各タッチの前半 1 話から 5 話までのオリジナルのセリフとシソーラスを用いたデータ拡張によって生成されたセリフを用いて, 学習時には 20% を検証用データとした．また, 評価用モデルは検証用データにおける正例の F 値が最大となるエポックでのパラメータを採用した．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．そして, 各タッチに対してモデルを作成し評価をした．表 5.2 に各実験で用いた正例と負例のデータ数を示す．

表 5.1: オリジナルデータ数

感情ラベル	ギャグ	少女	少年	青年	萌え	合計
喜楽	25	77	27	33	47	209 (31.7%)
ニュートラル	43	8	55	33	30	169 (25.6%)
驚愕	19	16	17	29	20	101 (15.3%)
悲哀	25	12	13	16	13	79 (12.0%)
恐怖	6	11	8	8	9	42 (6.4%)
憤怒	4	5	2	7	2	20 (3.0%)
嫌悪	2	4	3	3	4	16 (2.4%)
UNK	7	3	5	2	6	23 (3.5%)
合計	131	136	130	131	131	659

表 5.2: 実験で用いたデータ数

	感情ラベル	ギャグ	少女	少年	青年	萌え	合計
訓練用データ	喜楽	1115	2672	940	999	1766	7492 (37.1%)
	その他	2766	1396	3077	3146	2324	12709 (62.9%)
評価用データ	喜楽	10	38	12	14	22	96 (29.5%)
	その他	56	29	52	51	42	230 (70.5%)

5.3　実験 1：セリフ 1 文のみを入力とする感情推定

実験 1 では, セリフ 1 文のみを入力とする感情推定を行った. まず, セリフ 1 文を BERT に入力し 768 次元のセリフベクトルを得る. それを識別器への入力として, “喜楽” か “その他” かを推定した. 識別器としては 3 層 MLP を用いた.

本研究では, 各実験において Image Embedding 層には illustration2vec の筆者らによって公開されている事前学習済みモデルを用いた. Text Embedding 層には 2 種類の事前学習済み BERT モデルを用いた.

一つは, 京都大学が公開している日本語 Wikipedia より全 1800 万文を用いて事前学習させたモデル [24] (以下, “京大 BERT” と呼ぶ) , もう一つは hottolink 社が公開している大規模日本語 SNS コーパスを用いて事前学習させたモデル, hottoSNS-BERT [25] を用いた. BERT モデルは最終層のみを fine-tuning した.

表 5.3 に 3 層 MLP で用いたパラメータ, そして表 5.4 に学習で用いたパラメータを示す. 学習率は Optuna [26] によって最適なパラメータを探索した. また, 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を扱うことから, このまま学習すると上手く学習が進まないという問題がある. 本研究ではこの解決策として, 損失関数のクラス重みとして, 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. これにより, 正例がより正確に学習されると期待される. 表 5.5 に損失関数におけるタッチごとのクラス重みを示す.

表 5.3: 各実験における MLP のパラメータ

パラメータ	実験 1　実験 2
入力層次元	768　4864
隠れ層次元	30
出力層次元	2
活性化関数	tanh
ドロップアウト率	0.5

表 5.4: 学習パラメータ

パラメータ	実験 1・2
エポック数	50
バッチサイズ	16
損失関数	Cross Entropy Loss
最適化手法	Adam

表 5.5: 損失関数におけるクラス重み

タッチ	正例	負例
ギャグタッチ	0.773	0.227
少女漫画タッチ	0.272	0.728
少年漫画タッチ	0.831	0.169
青年漫画タッチ	0.824	0.176
萌えタッチ	0.594	0.406

5.4　実験 1：結果

表 5.6 に評価用データに対する実験 1 の結果を示す．表 5.6 における P-Recall, P-F 値はそれぞれ正例の再現率, F 値を表し, Acc は全体の精度を表している．また, 5 タッチ総合の結果は各タッチの結果における混合行列の和から算出している．本研究では, ベースラインは前述のように, 正例が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, すべての予測値が負例の場合を設定した．図 5.1 に Text Embedding 層に hottoSNS-BERT を用いた場合のギャグタッチにおける学習曲線を示す．上の図から Acc, P-F 値, loss を表し, 青線は訓練用データ, オレンジの線は検証用データを表す．

表 5.6 より, Acc に関してはどちらの BERT モデルを用いてもベースラインを超えた．また, ほとんどの評価指標において京大 BERT よりも hottoSNS-BERT の方が優れた結果を示した．図 5.1 において, 訓練用データと検証用データの学習曲線の推移が酷似しているのは, データの切り分け方に問題があると推測される．前述のように, 4 コマ漫画ストーリーデータセットには現状, 同一のストーリーを原則的に 4 コマ目がオチとなる “一般” と逆に 1 コマ目にオチを持ってきた “出オチ” の 2 種類の構造のみからなっており, セリフやその感情ラベルに大きな差はない．つまり, 似たセリフデータが 2 つずつ存在している場合が多い．そのため, 検証用データをランダムにサンプリングしたことでデータの分布にも差が生まれず, 訓練用データと検証用データの学習曲線の推移が酷似したと考えられる．

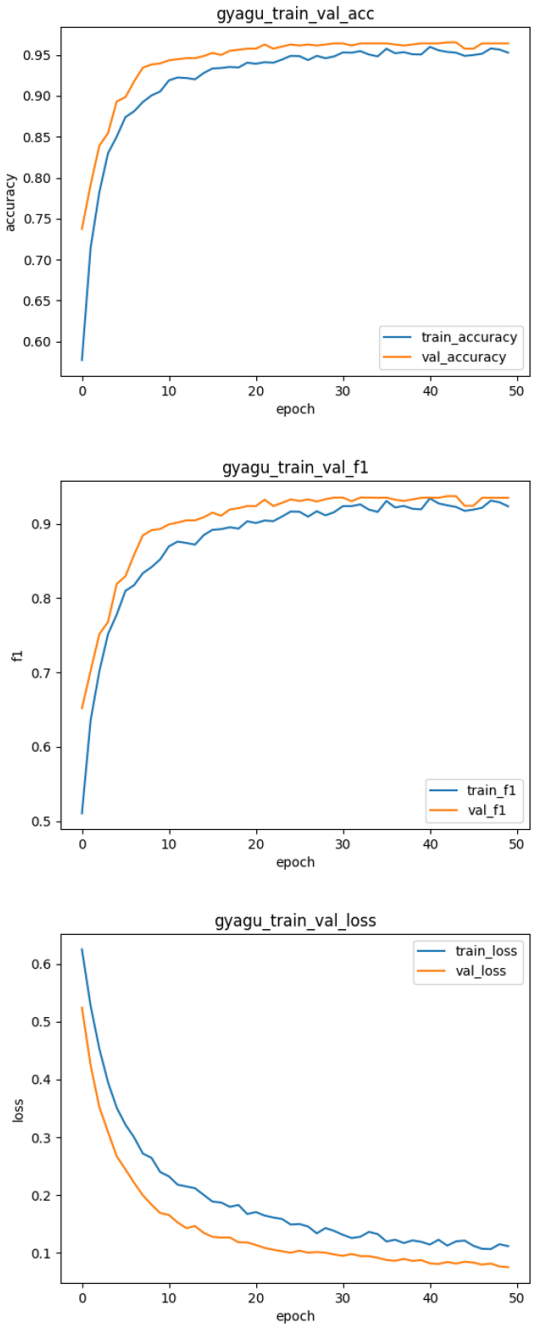

図 5.1: 実験 1 における学習曲線の例

表 5.6: 評価用データに対する実験 1 の結果

	ギャグタッチ			少女漫画タッチ		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
京大 BERT	0.818	0.200	0.250	0.612	0.711	0.675
hottoSNS-BERT	0.818	0.300	0.333	0.627	0.816	0.713
ベースライン	0.848	-	-	0.432	-	-

	少年漫画タッチ			青年漫画タッチ		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
京大 BERT	0.766	0.083	0.118	0.862	0.500	0.609
hottoSNS-BERT	0.813	0.167	0.250	0.846	0.643	0.643
ベースライン	0.812	-	-	0.784	-	-

	萌えタッチ			5 タッチ総合		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
京大 BERT	0.609	0.591	0.510	0.733	0.521	0.535
hottoSNS-BERT	0.688	0.500	0.524	0.758	0.583	0.586
ベースライン	0.656	-	-	0.705	-	-

5.5　実験 2：マルチモーダルな感情推定の検討

実験 2 では, 提案手法に則って, マルチモーダルな感情推定の検討を行った. 実験 1 と同様に 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単純に結合させた 4864 次元のベクトルを識別器である 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.

表 5.3 に 3 層 MLP で用いたパラメータ, そして表 5.4 に学習で用いたパラメータを示す. その他の実験設定は実験 1 と同様にした.

5.6　実験 2：結果

表 5.7 に評価用データに対する実験 2 の結果を示す. また, 図 5.3 に Text Embedding 層に hottoSNS-BERT を用いた場合のギャグタッチにおける学習曲線を示す. 実験 1 と同様に上の図から Acc, P-F 値, loss を表し, 青線は訓練用データ, オレンジの線は検証用データを表す. また, 5 タッチ総合の結果は各タッチの結果における混合行列の和から算出している.

表 5.7 より, Acc に関してはどちらの BERT モデルを用いても同様にベースラインを超えた. ほとんどの評価指標において京大 BERT よりも hottoSNS-BERT の方が上回り, 実験 1 および 実験 2 を通してそれぞれ最も高い値を示したことから, 京大 BERT よりも hottoSNS-BERT の方が漫画のセリフの分散表現の獲得手法として有効であると定量的に確認した. そして, 5 タッチ総合の結果より実験 1 よりも正例の識別率も向上したことから, マルチモーダルな感情推定の有効性を定量的に確認した. 図 5.3 において, 訓練用データと検証用データの学習曲線の推移が実験 1 の学習曲線より急勾配で早期に収束している. これは単純に識別器への入力次元数が増えたことに起因すると考えられる. そして, 多くのタッチにおいて実験 1 ではほとんど識別できていなかった正例データが識別できていたことを確認した. 図 5.2 にこれらの正例データを示す. 図 5.2 のオレンジの枠で囲われたセリフより, 文字だけでは人間の目で見ても判断が難しいが, 画像情報 (発話者の顔の表情) を加味することで喜楽寄りのセリフであると理解できるデータであることが分かる.

表 5.7: 評価用データに対する実験 2 の結果

	ギャグタッチ			少女漫画タッチ		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
京大 BERT	0.773	0.200	0.211	0.687	0.763	0.734
hottoSNS-BERT	0.818	0.400	0.400	0.761	0.816	0.795
ベースライン	0.848	-	-	0.432	-	-

	少年漫画タッチ			青年漫画タッチ		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
京大 BERT	0.703	0.417	0.345	0.769	0.643	0.545
hottoSNS-BERT	0.781	0.500	0.462	0.815	0.500	0.538
ベースライン	0.812	-	-	0.784	-	-

	萌えタッチ			5 タッチ総合		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
京大 BERT	0.641	0.500	0.489	0.715	0.583	0.546
hottoSNS-BERT	0.703	0.545	0.558	0.776	0.625	0.622
ベースライン	0.656	-	-	0.705	-	-

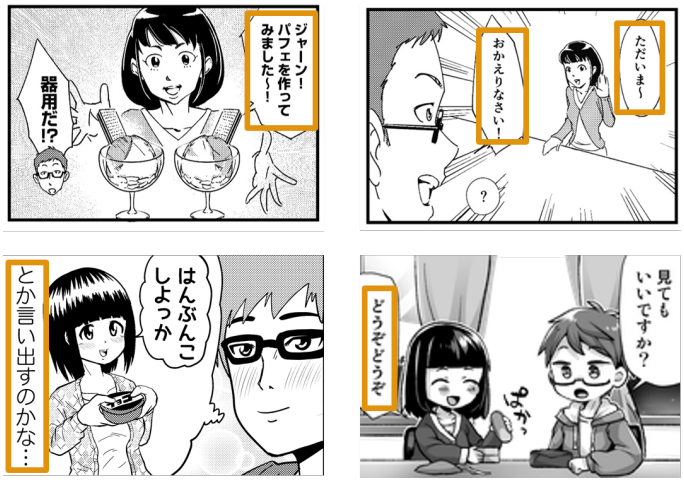

図 5.2: マルチモーダル化の精度への影響

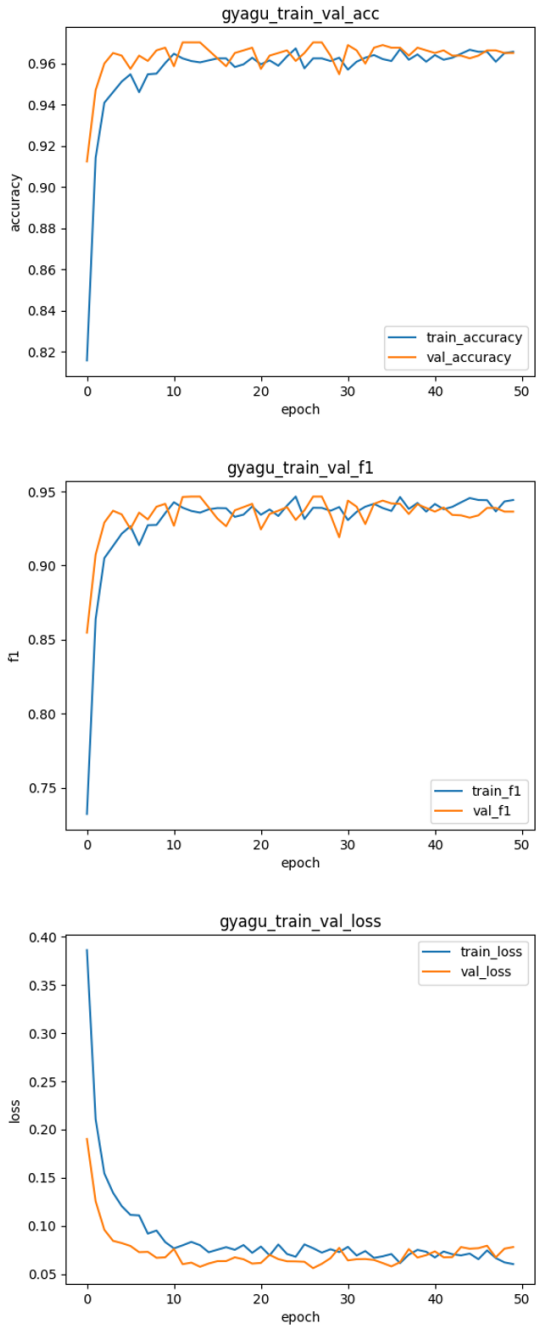

図 5.3: 実験 2 における学習曲線の例

5.7　考察

5.7.1　BERT モデルの合理性とシソーラスを用いたデータ拡張の問題

実験結果から, 定量的に京大 BERT よりも hottoSNS-BERT の方が漫画のセリフの分散表現の獲得手法として優れていることが示せた. ここで, 別の視点から京大 BERT と hottoSNS-BERT のどちらがより合理的な分散表現を得られているかについて考察する.

図 5.4 にシソーラスを用いたデータ拡張によって生成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を, 京大 BERT と hottoSNS-BERT を用いて計算したヒストグラムを示す. 横軸はコサイン類似度, 縦軸は区間内のデータ数を表している. また, 青色が hottoSNS-BERT, オレンジ色が京大 BERT を表す. ここで, コサイン類似度とは 2 本のベクトルがどれくらい同じ向きを向いているのかを表す指標であり, −1 から 1 までの値を取る. まったく同じ文章同士であればコサイン類似度は 1 であり, 似た文章であるほど 1 に近い値を示す.

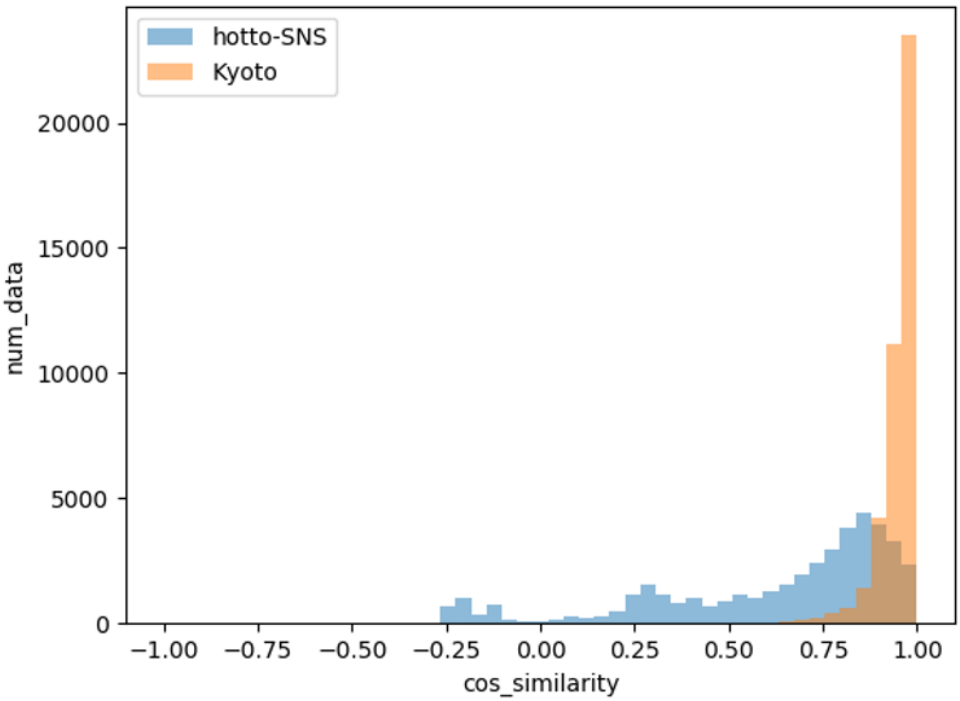

図 5.4: 拡張されたセリフと対応するオリジナルのセリフとのコサイン類似度

図 5.4 より, 一見して京大 BERT の方がコサイン類似度が高い区間に集まっていることから, 優れているように考えられる. しかし, これはデータ拡張の手法に大きな問題が無いという前提でのみ成立する. 本研究におけるシソーラスを用いたデータ拡張には, 文法的意味に齟齬が発生しているという問題がある. 例としては, 意味は通じるが形容詞であったものが名詞に置き換わっていたり, 文法的に間違ったものも含まれていた.

表 5.8 にシソーラスを用いたデータ拡張によって文法的齟齬が生じているデータに対するコサイン類似度を, 京大 BERT と hottoSNS-BERT から出力されたセリフベクトルを用いて計算した結果を示す. 表 5.8 より, 人間の目で見てもオリジナルのセリフから意味が遠いように思えたり, 文法的齟齬があるセリフに対して, 京大 BERT は高いコサイン類似度を示している一方で, hottoSNS-BERT は 0.8 付近にある一番大きな山には属さず, マイナスの値を取っている. このことから, 定性的にも京大 BERT よりも hottoSNS-BERT の方が漫画のセリフの分散表現の獲得手法として優れていると推測できる.

表 5.8: シソーラスを用いたデータ拡張による文法的齟齬

オリジナルのセリフ	拡張後のセリフ	京大 BERT	hottoSNS-BERT
去年は私が着たやつ	去年は私が着た若者	0.97	-0.19
そうですか	正しくですか	0.89	-0.19
僕はいいですけど、 気をつけてくださいね	僕はグーですけど、 気をつけてくださいね	0.96	-0.16
	僕はいいですけど、 真性をつけてくださいね	0.94	0.25

また, シソーラスを用いたデータ拡張による文法的齟齬に関する問題の解決策としては, 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.

そして, hottoSNS-BERT には特殊なトークンとして “<mention>” が存在する. これは, 事前学習で用いた日本語 SNS コーパスの中で, リプライ (返信) に該当するデータの文頭に付与されるトークンである. 図 5.5 に <mention> トークンの例を示す. 人手によってアノテートする必要はあるものの, このトークンを付与することによって, より会話の流れを汲んだ学習が可能になると考えられる. 応用例としては, 会話破綻の検出や人工知能を用いた会話の自動生成などが考えられる.

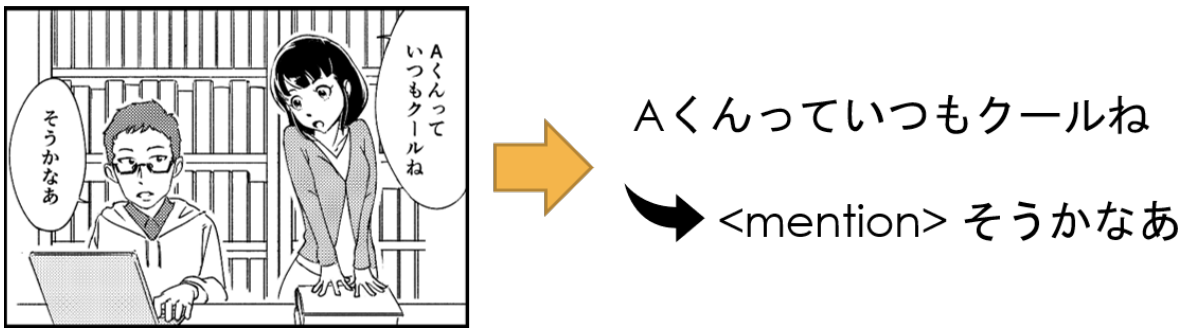

図 5.5: <mention> トークンの概要

5.7.2　illustration2vec モデルの合理性について

次に, 本研究で用いた illustration2vec モデルの合理性について考察する. 実験結果よりマルチモーダルな感情推定の有効性を定量的に確認した. しかし, どのような情報が取れているのかについては不明瞭である. そこで, t-SNE (t-distribution Stochastic Neighbor Embedding) [27] を用いてコマベクトルを 2 次元に圧縮し, 可視化した. 図 5.6 にその結果を示す.

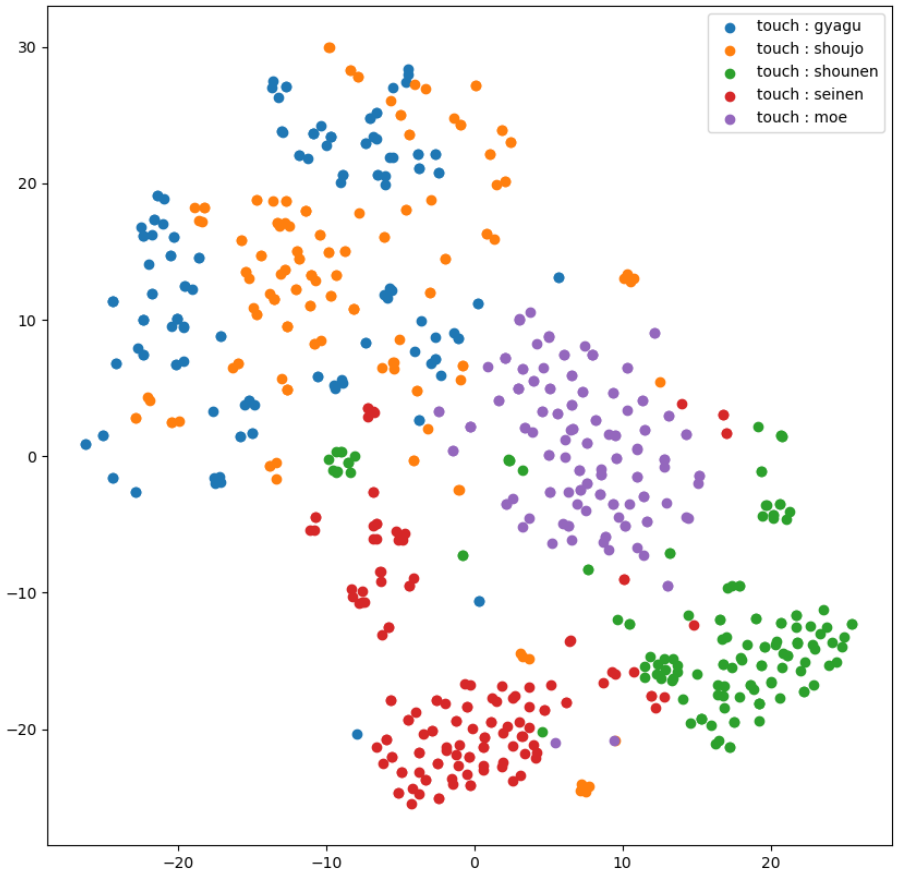

図 5.6: コマベクトルの t-SNE による次元圧縮結果

図 5.6 より, 少年漫画タッチ, 青年漫画タッチ, 萌えタッチではタッチごとに大きくまとまっていることが確認できる. ギャグタッチと少女漫画タッチについてもいくつかのまとまりが確認できることから, illustration2vec によって得られたコマベクトルはタッチの情報をうまく反映していると言える.

また, 本研究では illustration2vec をコマベクトルを得るためのエンコーダとしてのみ用いたが, illustration2vec では多クラスのラベル予測が可能である. 図 5.7 にその例を示す. ラベル予測の結果はラベル名とその確率で与えられる. 図 5.7 より, “モノクローム”, “1 人の女の子”, “1 人の男の子”, “眼鏡”, “ショートヘア”, といったようにコマ画像に含まれるものをよくとらえていることが分かる. このことから, より精度の高いマルチモーダルな感情推定を行うためには illustration2vec モデルの fine-tuning が有効であると考えられる. illustration2vec のラベルには “smile”, “angry”, “surprised”, といった感情に関係するものも含まれている. よって, 1 コマの中にどのような感情が含まれているか, といったメタ情報を付加し学習させることによって, より合理的なコマベクトルを得られると考えられる. また, 漫画では一般に読者にキャラクタの感情をより分かりやすく説明するための方法として, オノマトペやスクリーントーンの活用, 効果線, 吹き出しの形状などが挙げられる. 実際に, 上野らの研究 [16] によってキャラクタ間の感情をキャラクタの表情と吹き出しの形状から理解することが有用である可能性が高いことが示されていることから, コマ画像の特徴的要素を付加したマルチモーダルな感情推定についても有用であると考えられる.

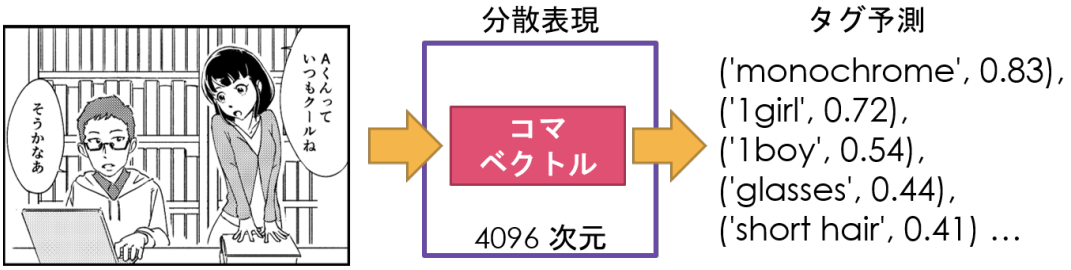

図 5.7: illustration2vec による多クラスのラベル予測例

5.7.3　マルチモーダル特徴量の結合方法の問題

本研究では, セリフベクトルとコマベクトルを単純に結合させることでマルチモーダルな感情推定を行った. しかし, それが最も良い方法であるという確証はない. この問題の解決策として, Differentiable Architecture Search (DARTS) [28] をはじめとしたニューラルネットワークの設計を自動化する手法, Neural Architecture Search (NAS) を用いて最適なマルチモーダル特徴量の結合構造を探索することなどが考えられる. こちらの適用については今後の課題とする.

5.8　補足 : 人手による新たなセリフデータの作成

本研究で行った実験には使用しなかったが, 4 コマ漫画ストーリーデータセットの問題点であるデータの少なさを改善するために, 人手による新たなセリフデータの作成を現在行っている. 以下にセリフデータの作成手順を示す.

1. データを拡張する対象のタッチを萌えタッチのみとし, また一部の 4 コマを除外する.
2. 使用するデータのセリフの吹き出しを白抜きし, 順番にセリフに対して固有の id を与える.
3. id に加え, 発話者の情報などをまとめた csv ファイルを用意し, アノテータに各 id に対応するセリフとその感情ラベルを付与してもらう.

また, 4 コマ漫画ストーリーデータセットの大きな特徴である “作者によるセリフの感情アノテーション” を担保するために以下の条件を付け加えた.

- 1 つの 4 コマ内での物語の一貫性を保つ.
- 登場人物の名前を独自に付ける.
  (オリジナルデータでは “A くん”, “B さん” といった抽象的な名前が付けられている.)

現状は自分を含めた 2 人分のアノテートが完了している. これらを拡張萌えタッチ A, 拡張萌えタッチ B とする. 図 5.8 に白抜きされた 4 コマ画像の例, 表 5.9 に, 図 5.8 に対応する拡張萌えタッチ A のデータ例を示す.

図 5.8: アノテーションのための白抜き 4 コマ画像の例

表 5.9: アノテートされたセリフと感情ラベルの例

id	セリフ (拡張萌えタッチ A)	感情ラベル
0	小池君 今週末の予定は？	喜楽
1	研究の進捗を生みますよ	ニュートラル
2	本当かなぁ？	ニュートラル
3	百合子さん こんにちは	ニュートラル
4	週末は随分と楽しんだみたいね	嫌悪
5	研究しないと留年するよ？	悲哀
6	分かってはいるのですが …	悲哀
7	一緒に卒業しようね！	喜楽

また, 表 5.10 に拡張萌えタッチ A, 拡張萌えタッチ B それぞれの感情ラベルごとのデータ数を示す. また, アノテートされたセリフの感情ラベルが一致した割合は 44.4% であった. この結果と表 5.10 から, この手法を用いることで多様な 4 コマ漫画ストーリーデータセットの拡張が可能であると言える. これらのデータを活用することでより汎化性能の高いモデルの生成が期待できる.

表 5.10: アノテートされたデータ数

感情ラベル	拡張萌えタッチ A	拡張萌えタッチ B
喜楽	33	27
ニュートラル	26	33
驚愕	11	11
悲哀	18	6
恐怖	1	10
憤怒	3	11
嫌悪	7	1
合計	99	99

6　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を推定した. 実験結果から口語的なコーパスを用いて事前学習された hottoSNS-BERT の方が漫画のセリフのより合理的な分散表現が得られることを確認した. また, コマの画像情報も考慮したマルチモーダルな感情推定手法を提案し, 実験的にその有効性を確認した. しかし, データの扱い方やネットワークの構築について更なる工夫が必要であると分かった.

今後の課題としては, 以下の点が挙げられる.

- コマベクトルの fine-tuning
- コマ画像の特徴的要素を付加したマルチモーダルな感情推定
- 最適なマルチモーダル特徴量の結合構造の探索
- 4 コマ漫画ストーリーデータセットの更なる拡張
- Manga109 やその他データセットを併用した半教師あり学習

謝辞

本研究を進めるにあたり御指導, 御鞭撻を賜りました森直樹教授に深く感謝申し上げます. 直接御指導頂きました岡田真助教には, 研究のアイデアや方針だけでなく, 論文の書き方や発表の作法に関することなど, 日頃から多岐に渡る御助言を頂きました. 心より御礼申し上げます. 最後に, 研究に関して建設的な意見をしてくださった諸先輩方, データセット作成に尽力してくださった方々, そして共に支え合いながら研究に取り組んできた同期の皆さんに感謝致します.

2021 年 2 月 26 日

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008. Curran Associates, Inc., 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] 松下光範. コミック工学：マンガを計算可能にする試み. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 29, pp. 199–199, 2013.

[5] 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*, Vol. 76, No. 20, pp. 21811–21838, 2017.

[6] 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.

[7] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.

[8] 京都大学大学院情報学研究科黒橋・河原研究室. 日本語形態素解析システム juman++ version 1.0. 2016.

[9] Greg Corrado Tomas Mikolov, Kai Chen and Jeffrey Dean. Efficient estimation of word rep-resentations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. *CoRR*, Vol. abs/1405.4053, , 2014.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[12] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[15] Masaki Saito and Yusuke Matsui. Illustration2vec: A semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*, SA '15, New York, NY, USA, 2015. Association for Computing Machinery.

[16] 上野未貴, 森直樹, 松本啓之亮. 漫画内の特徴的要素が与えるストーリーの印象についての検討. 第 30 回人工知能学会全国大会発表論文集, pp. 2J5–OS–08b–4in2, 2016.

[17] 上野未貴, 末長寿規, 井佐原均. 漫画中の表現獲得方法に基づくストーリー理解過程の解析. 第 31 回人工知能学会全国大会発表論文集, pp. 4F1–5in2, 2017.

[18] 迎山和司. 既存の漫画に基づいた 2 コマ漫画の自律生成. 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集, pp. 1K1OS2a02–1K1OS2a02. 一般社団法人 人工知能学会, 2018.

[19] Miki Ueno. Computational interpretation of comic scenes. In *Advances in Intelligent Systems and Computing*, Vol. 474, pp. 387–393, 2016.

[20] 野村俊太, 荒井幸代. 進化計算を用いた人間の感性理解のための遺伝子解析法. 第 31 回人工知能学会全国大会発表論文集, pp. 3H2–OS–04b–2, 2017.

[21] Miki Ueno, Naoki Mori, Toshinori Suenaga, and Hitoshi Isahara. Estimation of structure of four-scene comics by convolutional neural networks. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, p. 9. ACM, 2016.

[22] 藤野紗耶, 森直樹, 松本啓之亮. 3 分岐型畳み込みニューラルネットワークによる 4 コマ漫画の順序識別. 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集, pp. 1K1OS2a05–1K1OS2a05. 一般社団法人 人工知能学会, 2018.

[23] 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.

[24] 黒橋 禎夫柴田 知秀. Bert による日本語構文解析の精度向上. 言語処理学会 第 25 回年次大会, pp. pp.205–208, 2019.

[25] Sakaki, Takeshi, Sakae Mizuki, Naoyuki Gunji. Bert pre-trained model trained on large-scale japanese social media corpus. 2019.

[26] 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.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605, 2008.

[28] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. Vol. abs/1806.09055, , 2018.