文脈を考慮した漫画の内容理解アーキテクチャの提案

坪内温暉 [1]　鶴岡慶雅 [1]
[1] 東京大学
{tsubouchi,tsuruoka}@logos.t.u-tokyo.ac.jp

概要

漫画は画像とテキストが統合されたマルチモーダルの媒体である．その物語を理解するためにはコマ間の文脈的な関係性を捉えることが不可欠である．本研究では，動画像認識における空間・時間情報の分離処理に着想を得て，コマ内の要素間の関係とコマ間の文脈遷移を階層的に学習する漫画エンコーダのアーキテクチャを提案する．具体的には，ComicBERT の入力表現を拡張し，Spatial Encoder と Temporal Encoder の 2 つのモジュールを直列に配置することで，計算の効率化と文脈理解能力の向上を図る．次パネル予測タスクにおける評価実験の結果，提案手法は従来手法を上回る性能を示した．

1　はじめに

漫画は，視覚情報である画像と言語情報であるテキストが統合されたマルチモーダルな媒体である．その物語はコマ割りや吹き出しによるセリフ表現，オノマトペ，効果線など漫画特有の技法によって読者に伝えられる．

こうした特異性から，近年では漫画の解析や理解そのものを対象とした研究が，独立した分野として進められている．具体的には，コマ，キャラクター，文字といった漫画の構成要素を検出するモデル [1] やテキストに対する OCR を行うモデル，あるいは特定のコマやページに関する質問に回答するモデル [2] などが提案されている．しかし，既存の多くのモデルは解析や理解を行う対象が単一のコマ，あるいはページに限定されている．物語の流れを理解するためには，個々のコマやその構成要素の解析にとどまらず，コマ同士の文脈的な関係性も捉えるモデルを構築し，多数のコマにまたがる大域的な関係性を学習させることが必要不可欠である．

本研究では，漫画の物語の流れをより豊かに理解し，下流タスクに応用可能な漫画理解エンコーダを構築することを目的とする．具体的には，動画像認識の分野で用いられる手法に着想を得て，動画像認識モデルが空間方向の情報と時間方向の情報を分けて処理するという考え方を漫画理解に応用する．すなわち，コマ内の構成要素の関係性とコマ同士の文脈的な関連性の学習プロセスを分離したアーキテクチャを，ComicBERT [3] の手法を拡張することで構築する．提案するモデルの有効性を検証するため，次パネル予測 (Next Panel Prediction, NPP) というタスクを用いた評価実験を行った．実験の結果，提案手法が従来手法に比べ高い正答率を示した．このことは，動画像認識で用いられるような空間情報と時間軸情報の分離が，漫画の文脈理解にも応用可能であることを示唆する．

2　関連研究

2.1　ComicBERT

漫画に含まれるマルチモーダルな情報を統合的に扱い，構成要素同士の関係性を複数コマにわたって学習するフレームワークとして，ComicBERT [3] が提案されている．ComicBERT では，漫画を構成する要素としてコマ画像，キャラクター画像，吹き出し内のセリフ，ナレーションテキストの 4 つの異なるモダリティを定義している．各コマに対してこれらの要素を検出し，各モダリティについてその特徴量を抽出したのち，線形射影によって共通の埋め込み空間に写像することで異なるモダリティを統一的に扱うことができる．検出や特徴量の取得には，事前学習済みモデルが用いられている．得られた各モダリティの埋め込みは，コマ内の構成要素を表すトークン列としてまとめられる．それらをコマの読順に並び替えることで，Transformer [4] を基盤としたエンコーダである Comicsformer の入力として与えられる．この際，異なるコマを区別するため [SEP] トークンが挿入され，コマの境界が明示的に付与される．

モデルの事前学習には，masked Comic Modeling

(MCM) という自己教師あり学習タスクが用いられている．MCM ではモデルの入力として与えられるモダリティ特徴量の一部がマスクされ，周囲の情報からそれらを復元する．

このようにして学習されたモデルは，各構成要素同士の関係性を，複数コマに渡って学習され，クローズ形式のタスクをはじめとした複数の下流タスクで高い性能を示している．一方で，ComicBERT は一つのコマ内の要素同士の関係性と異なるコマの要素同士の関係性が単一の Transformer エンコーダ内で並列に処理される．この構造により，コマ内の要素の関係性とコマ間の関係性という性質の異なる 2 つの情報をモデルが十分に内包できていない可能性がある．これに対し本研究では，モダリティ定義や学習タスクは引き継ぎつつ，コマ内の要素間の関係性とコマ間の時間的・文脈的関係性を明示的に分離して処理するアーキテクチャを導入し，モデルがより適切にコマ間の関係性を理解することを目指す．

2.2 動画像エンコーダ

漫画はコマ画像がページ内に複数配置されるという特性をもっており，この構造はフレーム画像が時間軸方向に複数連なった動画の構造に類似している．そのため，動画像処理のアプローチが漫画にも応用できると考えられる．

動画像処理の分野では，時間軸方向にも情報をもつ動画データを効率的に処理するかが活発に研究されている．従来の 3 次元畳み込みニューラルネットワーク [5] に代わり，近年では Transformer を動画像向けに拡張したモデルが数多く提案されている．

ViViT [6] は，動画から抽出した時空間トークンを Transformer で処理するアーキテクチャである．特に，空間方向のエンコーダと時間方向のエンコーダを分離する Factorized Encoder という構成を用いることで，Attention 計算の負荷を軽減している．この設計は，各フレーム内の空間的特徴を抽出したのち，フレーム間の時間的関係を捉えるという処理フローを実現しており，計算速度と精度を両立させている．その他にも計算効率向上を目的として，時間方向の Attention と空間方向の Attention を交互に適用し学習を行う TimeSformer [7] や，局所的なウィンドウ単位での Attention 計算を行う Video Swin Transformer [8] など，Attention 計算に注目したモデルも提案されている．

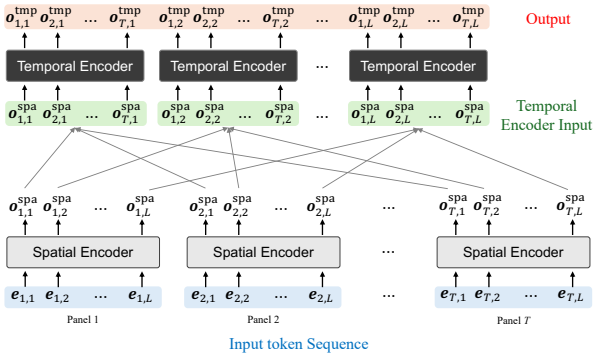

図 1　提案する階層的漫画エンコーダモデルの全体図

本研究では，このような既存の動画像エンコーダの構造的利点に着目し，コマ内の特徴量の抽出とコマ間の関係性の学習を分離するアプローチを漫画理解エンコーダに導入する．

3 提案手法

3.1 階層構造のエンコーダ

ViViT の Factorized Encoder にならい，漫画のエンコーディングにおいても図 1 のような階層的なエンコーダモデルを提案する．入力データは，T 個のコマからなる漫画のシーケンスである．各コマの埋め込み表現は L 個のトークンで構成される．アーキテクチャは空間エンコーダ (Spatial Encoder) と時間エンコーダ (Temporal Encoder) の 2 種類のモジュールからなり，まず各コマのトークン列を並列に Spatial Encoder に入力する．続いて，その出力テンソルを転置することで処理の主軸をコマ内からコマ間へと切り替え，Temporal Encoder に入力し，最終的な特徴量を得る．

このようにコマ内の関係性計算とコマ間の関係性計算を明示的に分割することで，計算効率を向上させながら，より豊かな複数コマ間の文脈情報をモデルが保持することができると考えられる．

3.2 コマの埋め込み表現

本節では，漫画の 1 つのコマ画像を L 個のトークン列に変換する手法について述べる．図 2 に，ComicBERT および提案手法における入力シーケンスの構成を示す．

ComicBERT では，漫画の構成要素として，以下の 4 つをトークンとして利用する．

- **PNL (Panel)**: コマ全体の画像特徴．
- **CHR (Character)**: キャラクターの画像特徴．

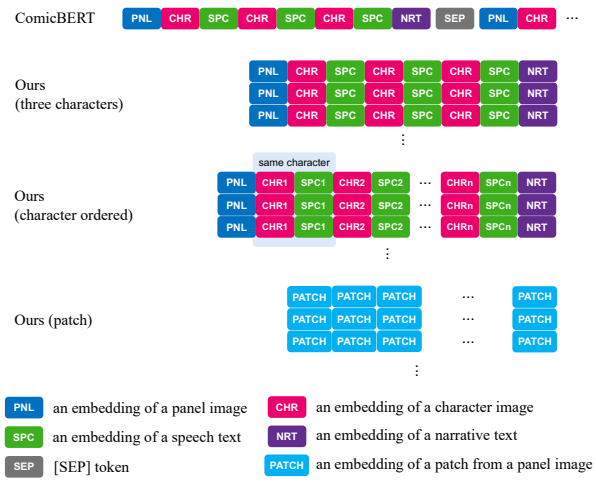

図 2　各手法でのコマの埋め込み表現

- **SPC (Speech)**: CHR トークンに埋め込まれたキャラクターの吹き出し内のセリフテキストの特徴．言語エンコーダにより埋め込まれる．
- **NRT (Narration)**: ナレーションテキストの特徴．言語エンコーダにより埋め込まれる．

これらの特徴量は，既存のエンコーダモデルにより抽出され，線形層を通して共通の埋め込み次元 D に射影される．さらに，各トークンの特徴量に対して対応するモダリティを表す埋め込みを加算し，コマの埋め込み表現としてモデルに入力する．

図 2 上段に示すように，ComicBERT では PNL トークンを 1 つ，CHR トークンと SPC トークンのペアを最大 3 組 (コマ内に 3 キャラクター以上存在しない場合やキャラクターにセリフがない場合は該当トークンはパディングされる)，NRT トークンを 1 つ並べコマの埋め込み表現としている．隣接するコマ同士は [SEP] トークンを挟んでトークン列を結合することにより区別している．一方，提案手法 (three characters) では，1 つのコマの埋め込み表現は ComicBERT と同じであるが，隣接するコマ同士は入力データの構造によって区別する．ComicBERT の入力テンソルの形状は (L, D) と表せるが，提案手法ではモデルのアーキテクチャに合わせて新たに軸を追加してコマの埋め込み表現を並べるため，(T, L, D) となる．また，CHR トークン数は最大 3 個という制限を取り除き，全てのコマを通して CHR トークンとして登場するキャラクターの順番を固定し配置する方法（Character Ordered）も導入する．このとき，例えば T コマのシーケンス内である特定のキャラクター A は各コマのトークン列内の同じ位置に登場するため，入力データがより構造化され，モデルの理解がより促進されることが期待される．

また，物体検出や埋め込み層のモジュールに依存しない別のアプローチとして，図 2 下段に示すパッチベース表現（Patch）も検討する．各コマ画像をいくつかのパッチに分割し，線形層に通すことでコマの埋め込み表現を得る．これは，T コマの画像シーケンスを T フレームからなる動画とみなし，従来の Video Transformer のように処理する手法である．本研究では，これらの入力表現それぞれについて学習を行い，下流タスクで評価を行い，その性能を比較検討する．

4　実験

4.1　実験設定

データセットには，Manga109 データセットおよびそのアノテーションデータ [9, 10, 11] を用いた．MCM による学習および下流タスクの学習，評価については，訓練・検証・テストデータの割合を 8 : 1 : 1 とし，全てのタスクについてこの分割方法は同一とした．これによって，下流タスクのテスト時にモデルが学習済みのコマのデータが出てこないようにした．

コマの構成要素の埋め込みについては，ComicBERT で用いられていたモデルと同様のものを使用した．具体的には，コマの全体画像の埋め込みには EfficientNet [12]，キャラクター画像の埋め込みには Character ReID model [13]，セリフとナレーションの埋め込みには Sentence Transformer [14] の DistilRoBERTa [15] を使用した．

MCM の学習時は，BERT [16] の Masked Language Model にならい，パディングではない入力トークンのうちの 20%をマスク対象とし，マスク対象となったトークンのうちの 80%を [MASK] トークンに，10%を別のランダムなトークンに，10%をそのままのトークンに変えてモデルに入力した．損失は，マスク対象のトークンに限定した最小二乗誤差とした．

4.2　評価タスク

本実験では，モデルが漫画の物語の文脈を適切に理解できているかを確認する評価タスクとして NPP を使用する．これは MangaUB [17] における next panel inference や ComicBERT における Scene-Cloze のようなタスクであり，コマのシーケンスが与えら

表 1　各手法での NPP 正答率 [%]

	easy	difficult
ComicBERT	72.06	64.88
Ours	**74.22**	**66.98**
Ours (character ordered)	69.83	65.35
Ours (patch)	61.74	61.20

れたときに，その次のコマが何かを選択肢の中から選ぶというものである．学習，推論時は各選択肢のコマについてその埋め込み表現を取得し，それまでのコマのシーケンスから得た埋め込み表現と結合させ，モデルに入力する．その後モデルの出力に対し Mean Pooling を施し，線形層を通して各選択肢につき 1 つの logit を得る．logit が最も高い選択肢をモデルの解答とし，その正答率をモデルの性能評価に用いた．難易度は easy と difficult の 2 種類を設定した．easy は選択肢のうち不正解のコマを正解のコマとは異なる本の中から抽出しており，difficult は不正解のコマを正解のコマと同じ本から抽出している．difficult は easy に比べ選択肢のコマの画風が似る傾向にあるため，より文脈そのものの理解が要求される．

MCM タスクで学習を行なったモデルに対し，NPP タスクの学習を行い，最終的にテストデータによる正答率の評価を行なった．学習の際，easy と difficult の比は 2 : 8 とし，損失関数としてクロスエントロピー関数を用いた．

本実験の詳細な設定は，付録 A に記載する．

4.3　結果

ComicBERT および提案手法における NPP の正答率を表 1 に示す．easy, difficult どちらの難易度においても，提案手法 (three characters) の正答率が従来手法である ComicBERT の正答率よりも高くなった．トークン列におけるキャラクター順を固定した手法 (character ordered) や，漫画のコマ画像の連続を動画のように扱う手法 (patch) については，どちらも 3 キャラクターに限定した手法 (three characters) と比較して正答率が低下した．

4.4　考察

本研究で提案した階層型アーキテクチャ (three characters) の高い正答率は，モデルがコマ間の時間的・文脈的関係性をコマ内の要素間の関係性と分離することによって複数コマにわたる漫画の文脈をより適切に捉えられたことに起因すると考えられる．特に difficult タスクでも高い正答率を示したことは，画風やキャラクターの外見的な特徴だけでなく，漫画の物語そのものを理解する能力が向上したことを示している．このように，提案した漫画エンコーダのアーキテクチャは，その構造的利点により，計算量の削減，文脈理解精度の向上の両方を達成したといえる．

一方で，トークン列内のキャラクター順を固定した手法 (character ordered) では，順序を固定しない手法よりも正答率が低下する結果となった．これは，入力トークン列の冗長化が主な要因であると考えられる．Character Ordered では，各コマでキャラクターの登場したかに関わらず，シーケンス内の全登場キャラクター分のトークンスロットを確保して配置を行うため，登場人物の少ないコマにおいてもトークン列が過度に長くなる傾向がある．その結果，入力情報が疎になり，学習の非効率化を招いている可能性が高い．本実験の結果は，順序を固定することによる入力テンソルの構造化という利点よりも，系列長の増大に伴う学習の非効率性や情報の希薄化といった欠点の影響が支配的であることを示唆している．これに対し，CHR トークン数を制限する手法は，系列長を抑制し，冗長なトークンを削減できるという点で有効であり，より効率的な文脈学習に寄与していると考えられる．

また，漫画のコマ画像をパッチ分割して入力する手法 (patch) での正答率も低くなった．パッチベースの手法では，画像のみの情報からキャラクターやテキストなどの漫画特有のモダリティの意味や関係性を学習する必要があるが，本実験で用いたモデルの規模ではそれらが十分に学習されなかったと推察される．

5　おわりに

本研究では，コマ内の関係性とコマ間の関係性を分離して処理する漫画エンコーダのアーキテクチャを提案し，NPP タスクにおいてその構造的な利点を示した．今後の課題として，複数の下流タスクにおける性能評価を通じたモデルの汎用性の検証や，1 コマあたりの CHR トークン数の上限の最適化，オノマトペなどの新たな漫画特有のモダリティの統合が挙げられる．

参考文献

[1] Ragav Sachdeva and Andrew Zisserman. From panels to prose: Generating literary narratives from comics. **arXiv preprint arXiv:2503.23344**, 2025.

[2] Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding. **arXiv preprint arXiv:2505.20298**, 2025.

[3] Gürkan Soykan, Deniz Yuret, and Tevfik Metin Sezgin. Comicbert: A transformer model and pre-training strategy for contextual understanding in comics. In Harold Mouchère and Anna Zhu, editors, **Document Analysis and Recognition – ICDAR 2024 Workshops**, pp. 257–281, Cham, 2024. Springer Nature Switzerland.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In **Proceedings of the IEEE international conference on computer vision**, pp. 4489–4497, 2015.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In **Proceedings of the IEEE/CVF international conference on computer vision**, pp. 6836–6846, 2021.

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In **Icml**, Vol. 2, p. 4, 2021.

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 3202–3211, 2022.

[9] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. **IEEE MultiMedia**, Vol. 27, No. 2, pp. 8–18, 2020.

[10] 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.

[11] Yingxuan Li, Kiyoharu Aizawa, and Yusuke Matsui. Manga109dialog: A large-scale dialogue dataset for comics speaker detection. In **Proceedings of the IEEE International Conference on Multimedia and Expo**, 2024.

[12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In **International conference on machine learning**, pp. 6105–6114. PMLR, 2019.

[13] Gürkan Soykan, Deniz Yuret, and Tevfik Metin Sezgin. Identity-aware semi-supervised learning for comic character re-identification. **arXiv preprint arXiv:2308.09096**, 2023.

[14] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, 2019.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)**, pp. 4171–4186, 2019.

[17] Hikaru Ikuta, Leslie Wöhler, and Kiyoharu Aizawa. Mangaub: A manga understanding benchmark for large multimodal models. **IEEE MultiMedia**, Vol. 32, No. 2, pp. 33–43, 2025.

A　実験設定の詳細

A.1　データセットの詳細

Manga109 は，日本の商業漫画全 109 作品からなるデータセットであり，総コマ数は 103,900 である．各ページに対してコマ，キャラクターの顔および身体領域，テキスト領域のバウンディングボックスの座標がアノテーションとして付与されている．また，キャラクターの場合はキャラクター ID，テキストの場合はその内容が付与されている．さらに，テキストとキャラクターを対応づけるアノテーションも提供されており，本研究ではこれらの情報を用いてモデルの入力となるトークン列を生成した．

A.2　前処理設定

入力として用いる各モダリティの特徴量について，線形層によって共通の次元に射影する前の次元を表 2 に示す．なお，Character の特徴量については，Character ReID model の backbone 埋め込み (2048 次元) とそれを identity head に通した埋め込み (256 次元) を結合している．

表 2　各モダリティの射影前の特徴量次元

Modality	Feature Dimension
Panel (PNL)	1280
Character (CHR)	2304 (2048 backbone + 256 identity)
Speech (SPC)	768
Narrative (NRT)	768

A.3　モデルのハイパーパラメータ

実験の際使用した，従来手法における Comicsformer と提案手法における Spatial Encoder，Temporal Encoder のハイパーパラメータを表 3 に示す．なお，patch size は提案手法 (patch) でのみ使用される．

表 3　モデルのハイパーパラメータ

Parameter	Comicsformer	Proposed Encoder
Embedding dimension	1024	
Number of heads	16	
FFN hidden size	4096	
Activation function	GELU	
Dropout rate	0.0	
Normalization	Pre-LN	
Number of layers	28	14 (Spatial) 14 (Temporal)
Patch size	–	18×18
Encoder structure	Single encoder	Spatial + Temporal

A.4　実験の詳細

MCM, NPP の学習時に使用したハイパーパラメータを表 4 に示す．

表 4　MCM および NPP タスク学習時のハイパーパラメータ

Parameter	MCM	NPP
Optimizer	AdamW	
Learning rate	1×10^{-4}	1×10^{-5}
LR Scheduler	Cosine Annealing	
Warmup ratio	0.1	
Weight decay	0.01	
Number of epochs	50	20
Batch size	16	8