D-022

内容ベース漫画推薦の実現に向けた条件付き画像類似度ネットワークの検討
Towards Content-Based Manga Recommendation via Conditional Similarity Networks

高橋 玲　伊藤 彰則　能勢 隆
Rei Takahashi　Akinori Ito　Takashi Nose

1　はじめに

推薦システムは，ユーザの嗜好に応じて商品やサービスを提示する技術である．漫画市場の拡大に伴い，漫画に特化した推薦技術への関心が高まっている．漫画推薦ではジャンルなどのメタ情報や，概要・レビューなどのテキスト情報に基づくアプローチが多く見られる一方，漫画の原稿画像を活用した推薦手法の研究は少ない．また，推薦根拠を明示できず，ユーザによる調整が困難な手法も依然として多い．本研究では，漫画の原稿画像に基づく透明性と制御性を備えた内容ベース推薦の実現を目指す．本稿ではその初期段階として，漫画の原稿画像を対象に，特定の条件（年代・対象・ジャンル）に基づく類似性を分離して学習する深層距離学習モデルの構築に取り組む．漫画推薦への応用を見据え，条件に応じて原稿画像の類似度を埋め込むモデルの構築に取り組む．

2　内容ベース推薦と深層距離学習

内容ベース推薦は，ユーザに依存せずアイテム自体の特徴に基づいて推薦を行う手法である．近年ではコンテンツの特徴に基づいた推薦を実現する手法として，深層距離学習（Deep Metric Learning）が注目されている．深層距離学習は埋め込み空間上において，基準となるデータ（Anchor）と同一のクラスラベルや条件を持つ正例（Positive）の距離を縮め，異なるクラスを持つ負例（Negative）を遠ざけるようにモデルを学習する手法である．データ間の類似度を空間上で直接学習できることから，内容ベース推薦への応用が進んでいる．

漫画推薦においても画像情報を活用した深層距離学習ベースの推薦手法が提案されている．白石らは表紙画像から得た特徴ベクトル間の類似度に基づく推薦手法を提案し [1]，渡邉らはキャラクターの服装に注目した推薦手法を示している [2]．これらは視覚的特徴を用いた推薦の有効性を示す一方，特徴空間上の位置関係に基づいて推薦が行われるため，推薦理由を人間の解釈可能ではないという課題も残されている．

3　条件付き漫画画像類似度ネットワーク

本稿では制御性・透明性の高い漫画推薦への応用を目的として図 1 に示す Conditional Similarity Networks（CSNs）[3] を用いた条件付き画像類似度ネットワークを提案する．CSNs はデータ間の類似性をジャンルなどの条件ごとに分離された特徴空間上で学習する．入力画像 x に対して画像認識モデルに基づく特徴抽出器 $g(x)$ により高次元の特徴ベクトルを得，条件 c に対応するマスクベクトル m_c を特徴ベクトル $g(x)$ にアダマール積として適用することで，条件ごとの類似性を反映した埋め込み表現 $z_c = m_c \odot g(x)$ を得る．これにより，どの基準に基づいているかが明確となり，推薦理由の透明性が向上する．また条件により変化する類似関係を独立した空間で表現することにより，多様な関係性を柔軟に表現できるという点で，性能の向上が期待される．

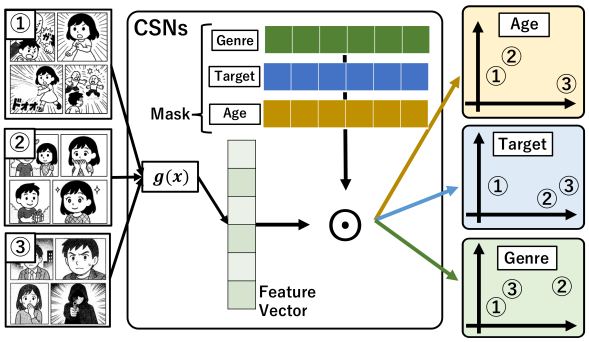

図 1: CSNs に基づく条件付き画像類似度ネットワーク

特徴抽出器には，CLIP により事前学習された ResNet-50（RN50）および Vision Transformer（ViT-B/32）[5] を導入する．本手法では，ジャンル（Genre），対象（Target），年代（Age）の 3 つの条件 c を考慮し，条件 c に対応したマスク m_c を学習する．これを特徴ベクトルに適用し選択的に活性化し，部分空間において距離学習を行う．条件 c に対するトリプレットの距離差 D_{tri} は，式 (1)，式 (2) のように定義される．

$$D_{\text{tri}}(x_i, x_j, x_l, c; m) = D(x_i, x_j; m_c) - D(x_i, x_l; m_c) + h \quad (1)$$

$$D(x_i, x_j; m_c) = \left\| g(x_i) \odot m_c - g(x_j) \odot m_c \right\|_2 \quad (2)$$

ここで，x_i は Anchor，x_l は正例（Positive），x_j は負例（Negative），c は条件，m は条件に対応するマスクベクトルを行方向に並べたマスク行列であり，m_c はそのうち条件 c に対応するマスクベクトルである．$D(x_i, x_j; m_c)$ は，マスク m_c を適用した特徴ベクトル間のユークリッド距離を表す．h はマージンであり，Anchor と正例の距離が負例よりも少なくとも h だけ小さくなるようにモデルに制約を課す．この距離差に基づき，トリプレット損失 $\mathcal{L}_T$ はヒンジ損失として式 (3) のように定義される．

$$\mathcal{L}_T(x_i, x_j, x_l, c; m) = \max\left\{0,\ D_{\text{tri}}(x_i, x_j, x_l, c; m)\right\} \quad (3)$$

このようにして，マスク m_c および特徴抽出器 $g(x)$ をトリプレット損失によって学習し，条件ごとの意味的部分空間において視点別の距離最小化を実現する．

4　実験

4.1　データセットおよびトリプレット構成

学術研究向けの漫画画像データセット Manga109[4] の 109 作品を，年代（Age），対象（Target），ジャンル（Genre）のラベル分布が偏らないように，Jensen-Shannon 距離に基づいて訓練・検証・テスト（70/15/24 作品）に分割した．各分割セットに対して，属性ラベル c に基づき，アンカー $w^{(a)}$，類似 $w^{(p)}$，非類似 $w^{(n)}$ からなるトリプレット $(w^{(a)}, w^{(n)}, w^{(p)}, c)$ を構成した．さらに，各作品からランダムに 1 ページを抽出し，ページ単位でのトリプレットを生成して学習に用いた．

表 1: トリプレット数別の各モデルの Total Accuracy（%）

# of Triplets	ResNet18	CLIP-RN50	CLIP-ViT
5k	52.94	56.17	51.40
12.5k	54.21	57.48	52.43
25k	54.52	**60.31**	53.28

表 2: 25k 件学習時の各モデルの条件別正解率（%）

Model	Age	Target	Genre	Total
ResNet18	51.71	62.04	49.84	54.52
CLIP-RN50	**58.02**	**66.57**	**56.04**	**60.31**
CLIP-ViT	51.07	58.25	50.44	53.28

4.2 訓練詳細

提案法の検証および評価は，埋め込み空間上でのトリプレット正解率（Accuracy）に基づいて行った．具体的には条件付きトリプレット $(w^{(a)}, w^{(n)}, w^{(p)}, c)$ に対し式 (1) における距離差が 0 以下となるかの正解率を算出した．トリプレットはランダムに構成し，チャンスレートは 50% である．評価に用いるモデルは，検証セット上で Accuracy が最も高かったエポックにおける重みを用いた．検証セットには学習に用いたトリプレットの約 10%を割り当てた．

比較対象として，CLIP ベースの RN50 および ViT-B/32 に加え，CSNs の元論文で用いられた ResNet-18 を用いた．各モデルには共通の 63 次元の埋め込み層を追加し，全層をファインチューニング可能とした．最適化には Adam[6] を用い，その他の設定は CSNs に準拠した．学習には条件 c を均等にサンプリングした 8 トリプレットからなるミニバッチを使用した．トリプレット数は条件 c ごとに 5k，12.5k，25k 件の 3 設定とし，テストには全モデル共通で各条件 4.8k 件のトリプレットを使用した．

4.3 トリプレット数の違いによる正解率の変化

表 1 に，トリプレット数を変化させた場合における各モデルの Total Accuracy の比較結果を示す．表から，トリプレット数の増加に伴い正解率が全体的に向上する傾向が確認できる．特に CLIP-RN50 では 25k 件のトリプレットを用いた際に 60.31%の Total Accuracy を記録し，最も高い性能を示した．一方 ResNet18 および CLIP-ViT における性能向上は緩やかであった．

CLIP-RN50 は多様な事前学習データに基づき，視覚特徴の表現力が高く，条件 c に応じた類似度の識別に有効であったと考えられる．一方で正解率は最大でも約 60%にとどまり，ImageNet ベースの CSNs 元論文の結果（約 90%）と比べて低く，漫画画像の視覚情報のみを用いた類似度学習には限界があることが示唆される．

4.4 条件別のトリプレット正解率

続いて各モデルの異なる条件（年齢層（Age），対象読者（Target），ジャンル（Genre））に対する分類性能を比較した．表 2 に，学習に 25k 件のトリプレットを用いた場合の，各モデルにおける（Total Accuracy）および条件別正解率（%）を示す．CLIP-RN50 は全条件において他モデルより高い性能を示し，Target 条件では 66.57%と最も高かった．これは，読者層に関わる構図や描線といった視覚的特徴を，視覚と言語の事前学習を通じて適切に捉えた結果と考えられる．

表 3: 25k 件学習時のマスク有無別 Total Accuracy（%）

Mask	RN18	CLIP-RN50	CLIP-ViT
w/o	55.39	**60.19**	54.02
w	54.52	**60.31**	53.28

一方，Genre 条件では 50%前後にとどまり，ジャンルのような抽象度の高い属性は画像特徴のみでは捉えにくいことが分かる．また 1 作品に対して 1 つのジャンルラベルしかないため，複数ジャンルの要素を含む作品に対して正確な学習が困難であった可能性もある．Age 条件では，CLIP-RN50 が 58.02%と最も高かった一方，ResNet18・CLIP-ViT は 51%前後にとどまった．この差異から，CLIP-RN50 が登場人物の描写や構図など，年代的傾向に関連する視覚要素を他モデルよりも効果的に学習していたと考えられる．

4.5 マスク有無別の正解率

本節では，マスクの有無が推薦精度に与える影響について検討する．表 3 は，各モデルに対して 25k 件のトリプレットで学習を行った際の，マスクあり（w）／なし（w/o）の Total Accuracy（%）を示している．CLIP-RN50 では，マスクありの条件においてわずかに正解率が向上したが，ResNet18 および CLIP-ViT ではいずれも精度が低下した．ただし，いずれのモデルにおいても差異は 2%未満であり，マスクの有無による性能への影響は限定的であった．この結果は，漫画画像においてジャンル，対象読者層，年代といった条件間の視覚的差異が曖昧であること，マスクの次元数に制約があること，および抽象的な条件と視覚特徴との対応関係が弱いことなどに起因すると考えられる．加えて，ジャンル・年代・対象読者層といった条件は，そもそも視覚的特徴ベースで明確に分離できるとは限らず，これらのカテゴリ自体に重なりや曖昧さが内在していた可能性もある．

5 おわりに

本研究では漫画推薦への応用を見据え，条件付き類似度学習に基づく画像特徴の扱いについて基礎的な検証を行った．CLIP ベースのモデルは比較的高い正解率を示した一方で，抽象的な条件に対しては視覚特徴のみでの判断に限界が見られた．今後はマスク構造の改良やテキスト情報との統合などを通じてより柔軟で解釈可能な推薦を目指す．また主観評価を通じて，実用上求められる精度水準の検討も進める．

参考文献

[1] 白石 et al.,“セレンディピティ誘発のための表紙を活用したコミック推薦システムの提案”, 情報処理学会全国大会講演論文集 2022, Vol. 2022, No. 1, pp. 553–554.

[2] 渡邉 et al. ,“印象にもとづくコミック検索に向けた服領域自動抽出と印象推定に関する検討”, コミック工学研究会予稿集 2020, pp. 38–45.

[3] Veit et al., “Conditional similarity networks”, in Proc. CVPR 2017, pp. 830–838.

[4] Fujimoto et al., “Manga109 dataset and creation of metadata”, in Proc. MANPU 2016 pp. 1–5.

[5] Radford et al.,“Learning transferable visual models from natural language supervision”, in Proc. ICML 2021, pp. 8748–8763.

[6] Diederik P. Kingma, Jimmy Ba,“Adam: A Method for Stochastic Optimization”, in Proc. ICLR 2015. pp. 1–13 .