ファッションアイテムの分散表現に基づくコーディネート理解

Understanding of Fashion Items Coordination based on distributed expression

林 美衣 *1　森 直樹 *1
Mie Hayashi　Naoki Mori

*1大阪府立大学
Osaka Prefecture University

Recently, analysis of the fashion outfits has become an attractive research topic in the artificial intelligence fields. However, understanding of fashion is not easy because of this problem including human creativity and Kansei. In this study, we have proposed a machine learning system to solve the problem of fashion outfits composing. In the proposed method, distributed representation by CAE is applied to fashion items, and deep learning techniques based on MLP and LSTM are utilized. The effectiveness of the proposed method is confirmed by computer simulations taking original four-choice questions for fashion items in the Polyvore dataset as examples.

1. はじめに

近年，機械学習の発展を背景とした人工知能 (Artificial Intelligence : AI) が注目を浴びている．AI は単純なパターン認識では人間の能力を凌駕する一方で，人間の感性, 特に創作に関する分野への AI の適用はいまだに難しく AI 研究の重要な課題とされている．

そこで本研究では，人間の感性の多様性や主観を多く含むため，工学的な処理が難しいとされてきたファッションに着目し，アイテム画像の分散表現化に基づくファッション理解について検討する．本研究では，まず複数の衣服やアクセサリーの画像で構成されたコーディネートのデータを用いて深層学習に基づく学習システムを構築し，次に衣服の組合せに関する問題を設定してコーディネートを理解する人工知能手法を提案する．

2. 要素技術

2.1 深層学習

本研究では，以下の深層学習 (Deep Learning) を用いた．

2.1.1 Convolutional Autoencoder

Autoencoder (AE) とはニューラルネットワークを用いて入力データの次元を削減し，有効な特徴ベクトルを獲得する手法である．AE は Encoder と Decoder の主に 2 つのパートからなっている．Encoder では，入力されたベクトルの次元圧縮をする．一方 Decoder では，圧縮されたベクトルから元の入力を再現しようとする．元の入力と出力が同じになるようにネットワークを学習させるため，潜在変数は入力のデータの重要な特徴量を効率的に保持した形になっている．

そして Convolutional Autoencoder (CAE) は AE に畳み込み層を追加した拡張手法で，画像から特徴量を抽出することに適している．

2.1.2 Long short-term memory

Long short-term memory (LSTM) [Gers 99] は，RNN [Zaremba 15] よりも長期間の時系列性を持つデータの利用に優れたモデルである [Hochreiter 97]．LSTM では中間層において，RNN におけるノードの代わりに入力値や重みを保持することができる LSTM-Block と呼ばれる構造を持っており，重み係数を 1 にすることで誤差の消失などを回避し過去の情報を保持することを可能としている．

LSTM-Block にはメモリセル，入力判断ゲート，忘却判断ゲート，出力判断ゲートから構成されている．メモリセルは過去の状態を記憶する役割を持つ．入力判断ゲートはメモリセルに加算される値の調整し，重要でない過去の情報がメモリセルの重要な情報を上書きしてしまうことを防ぐ．忘却判断ゲートはメモリセルの値が次の時刻で保持される割合を調整し，不要となった過去情報の破棄を実現する．出力判断ゲートはメモリセルの値が次層に与える影響の調整をする．この 3 つのゲートとメモリセルが LSTM-Block の大きな特徴となっている [岡谷 15].

連絡先: 林 美衣，大阪府立大学 工学研究科，大阪府堺市中区学園町 1-1，hayashi@ss.cs.osakafu-u.ac.jp

3. Polyvore Dataset

本章では，本研究で用いたデータセットである Polyvore Dataset [Han 17] について述べる．Polyvore とは EC サイトの服の画像を，ユーザがコラージュして投稿するサービスで，同サイトの投稿を収集したものが Polyvore Dataset である．Polyvore Dataset ではアイテムがおおよそ，トップス類，ボトムス類，靴，アクセサリの種類順になっている．そして，トップスから順に index が振られていく．このため index が大きくなるに従って，アイテム種別は大まかに定まるが，一意には対応していない．

このデータセットには 21,889 のコーディネートがある．1 つのコーディネートに含まれるアイテムの数は各コーディネートによって異なるため，本研究では index 1〜8 の計 8 アイテムを使用することとする．具体的には，アイテム数が 8 未満のデータは使用せず，アイテム数が 9 以上であるデータに関しては，index 9 以上のアイテムを削除して，index 1〜8 のアイテムのみ使用した．

4. 従来手法

ファッション理解については LSTM に基づく従来手法 [Han 17] が提案されている．同手法では，本研究と同様に Polyvore dataset を用いて bidirectional LSTM (Bi-LSTM) によるコーディネイトの学習を試みている．また，Task 1: 既存のファッションアイテムセットに適するアイテムの推薦, Task 2: キーワードまたは画像入力に基づくコーディネートの推薦,

Task 3: 入力コーディネートの評価の 3 問題を設定している．一方で，従来手法では画像そのものを畳み込みニューラルネットワーク (Convolutional Neural Network: CNN) の入力としており，個々のアイテムの分散表現化については検討されていない．

5. 提案手法

本研究では，ファッションアイテムを理解する手法を構築するため，以下の手法を提案する．

5.1 CAE に基づく分散表現化

提案手法では，まず従来手法では検討されていない CAE によるアイテム画像の分散表現化をする．従来研究のように既存の CNN モデルを用いる手法も考えられるが，ファッションアイテム固有の特徴に着目するためには CAE の方が適している場合が多い．また，CAE を用いる場合は，分散表現から画像を再現できるという利点もある．提案手法による分散表現は，服の色や形といった性質に基づいて作られた潜在空間中において，性質の類似性が高いアイテムの推薦に利用可能である．よって本研究では，適切な分散表現の獲得と，得られた分散表現の効率的な利用が必要となる．

5.2 問題 Question 1 および 2

コーディネートは明確な答えのない問題であるため，提案手法の性能を評価するために，独自の問題を提案する．以下に問題の詳細を示す．

コーディネート理解の確認のために，予測候補と選択肢の分散表現におけるユークリッド距離に基づく 4 択問題を設定する．4 択問題は テストデータの他のコーディネートから候補を 3 つ選び，正解のアイテムと合わせて 4 択としている．この 4 択問題の正解率 （accuracy) を評価指標とする．

また，不正解となる選択肢を他のコーディネートの index 1 から集めた場合と index 3 から集めた場合の 2 通りの問題を作成し，Question 1, Question 2 とする．Question 1 ではトップスやワンピースなど類似した衣服が選択肢を構成するため，問題としての難易度は高くなる．一方で Question 2 は，選択肢が複数のカテゴリのアイテムからなるため，難易度の低い問題となる．なおこの問題は 7 種のコーディネートアイテムに最も適したコーディネートアイテムを選ぶことに相当する．

5.3 深層学習に基づく学習器

本研究では与えられたコーディネートの分散表現を入力，適切なファッションアイテムの分散表現を出力として学習する手法を提案する．具体的な学習として，データセットのコーディネートを正例として，index 2〜8 の分散表現を学習器に入力し， index 1 の分散表現を出力させた．ここで今回は具体的な学習器として MLP と LSTM の 2 種類の深層学習モデルを使用した．MLP では複数の分散表現を連結して入力し，LSTM では index 順を 1 つの系列として解釈して学習するものとした．また Random Forest を用いて CAE に基づく分散表現の有効性を確認した．

6. 数値実験 1

6.1 CAE の学習

Polyvore Dataset から train data 135,180 枚, test data 33,795 枚を使用し，アイテム画像を CAE に学習させ，中間層からアイテム画像の分散表現を得た．validation data は train data の 2 割を使用した．また，CAE の学習終了条件として，validation loss を監視した early stopping を用いた．CAE の中間層の次元として 512, 1024 および 2048 の 3 種類を設定した．続いて CAE で得られた分散表現の有効性を確認するための問題を設定する．

6.1.1 問題設定

CAE の中間層から得た分散表現の有効性を確認するため，分散表現を入力とし，Random Forest で アイテムの index を識別する．表 1 に Random Forest のパラメータを示す．パラメータはそれぞれ Optuna [Akiba 19] で設定をした．

6.1.2 結果と考察

表 2 に 512, 1024 次元の分散表現それぞれを Random Forest で識別した結果とその混同行列を示す．

識別結果はどちらも 30% に満たないが，全ての出力を index1 とした場合のベースラインを超える精度が得られている．index によってアイテムのカテゴリが決まるわけではないので，混同行列のおおよそ真値付近に識別結果が集中していればよい．512 次元はベースラインよりも高い精度は出しているものの，真値付近にあまり分布していないことから，次元を圧縮しすぎてうまく特徴量を獲得できていないと考えられる．

一方で，1024 次元では精度が 512 次元のものよりも高く，混同行列の対角線上に多く分布しており，適切に特徴量の獲得ができたと考えられる．2048 次元の結果に関しても識別率は 29.9% であり，1024 次元の結果と同様な分布が得られた．

表 1: Random Forest のパラメータ

	512 次元	1024 次元	2048 次元
n estimators	100	30	80
max features	5	5	20
min samples split	12	12	8
max depth	8	56	24

表 2: Random Forest での識別結果の混同行列

baseline = 0.125									
512 次元		予測値							
識別率：0.249		1	2	3	4	5	6	7	8
真値	1	**3288**	332	89	517	112	84	11	3
	2	**2003**	**972**	290	**817**	129	88	14	7
	3	**1254**	678	428	**1630**	150	176	22	28
	4	**1254**	196	323	**1962**	309	303	50	45
	5	**1373**	203	127	**1497**	403	482	89	149
	6	**1338**	170	133	**1060**	441	542	164	311
	7	**1308**	179	102	**927**	381	487	228	468
	8	**1192**	124	103	738	299	427	203	583
1024 次元		予測値							
識別率：0.282		1	2	3	4	5	6	7	8
真値	1	**2555**	**839**	286	240	185	126	108	54
	2	**1191**	**1592**	683	380	215	152	124	56
	3	574	**868**	**1096**	**967**	425	224	157	82
	4	448	379	**881**	**1207**	718	415	220	125
	5	422	352	518	**952**	789	606	402	310
	6	437	274	343	618	699	681	608	560
	7	360	258	287	482	528	625	638	795
	8	300	187	245	362	403	525	713	**944**

7. 数値実験 2

提案する MLP および LSTM による分散表現の学習について数値実験をした．

7.0.1 MLP に基づく提案手法の構築

index 2〜8 のそれぞれの分散表現を連結した $512 \times 7 = 3,584$ 次元の 1 次元ベクトルを MLP に入力し，出力は 512

表 3: MLP のパラメータ

	512 次元	1024 次元	2048 次元
epoch	31	17	89
batch size	32	32	32
activation	ReLU	ReLU	ReLU
loss function	Mean Squared Error (MSE)	MSE	MSE
learning rate	0.00367	0.00236	0.00001
unit 1	9	13	247
drop out 1	0.458	0.394	0.460
unit 2	42	136	-
drop out 2	0.361	0.215	-

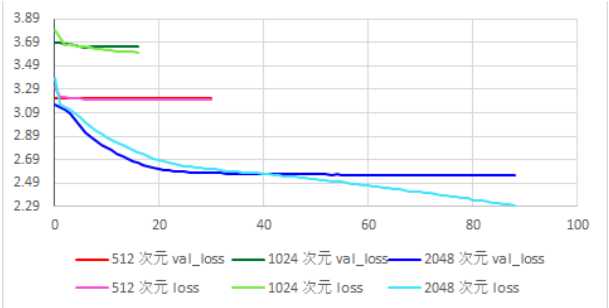

図 1: MLP 学習時の training loss と validation loss の推移

次元のベクトルとした．1024 次元, 2048 次元でも同様にベクトルを連結させ MLP で学習させた．

表 3 に用いた MLP のパラメータを示す．パラメータとモデルは Optuna を用いて調整した．また，CAE の学習終了条件として，validation loss を監視した early stopping を用いた．図 1 に，512 次元，1024 次元，2048 次元の MLP 学習の loss と validation loss の推移を示す．図の横軸は epoch 数を，縦軸は loss の値を示している．512 次元も 1024 次元も validation loss の減少が見られるので，学習が進んでいることがわかる．また減少度合から 2048 次元が最も学習が進んでいる．

7.0.2 LSTM に基づく提案手法の構築

index 2〜8 のファッションアイテムの分散表現を LSTM に入力し，index1 の分散表現を出力として学習させた．ここでは分散表現は 2048 次元のものを用いた．

表 4 に用いた LSTM のパラメータを示す．また，CAE の学習終了条件として，validation loss を監視した early stopping を用いた．図 2 に LSTM 学習時の training loss と validation loss の推移を示す．図の横軸は epoch 数を，縦軸は loss の値を示している．

7.1 結果と考察

7.1.1 MLP に基づく提案手法の結果

表 5 に提案手法で学習した時の 512 次元，1024 次元，2048 次元それぞれの accuracy を示す．Question 1 は 512 次元，1024 次元ともにベースラインを下回る結果となった．2048 次元ではわずかに上回る結果が出た．これは正解の選択肢と極めて類似したアイテムが選択肢に含まれており，不正解の回答でもコーディネートが成り立つ場合が多かったのだと推測される．Question 2 に対しては，50% 以上の精度が得られた．

index 1 の accuracy は低かったが，正解のアイテムが 4 択の中で何番目に予測アイテムとの距離が近いかが重要であると考え，1 番目に近ければ 1 を，2 番目に近ければ 0.5 を，3 番目に近ければ 0.25 を，4 番目に近ければ 0 を与え，それらの平均を

表 4: LSTM のパラメータ

epoch	84
batch size	32
loss function	MSE
learning rate	0.001
hidden layer size	256
optimizer	Adam

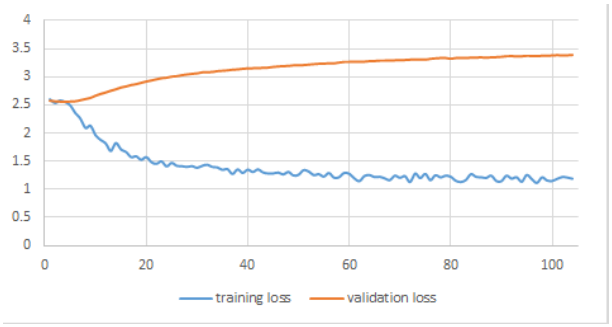

図 2: LSTM 学習時の training loss と validation loss の推移

出し，順位の指標になるものを計算した．表 6 に計算結果を示す．この場合のベースラインは $(1+0.5+0.25+0)\div 4 = 0.4375$ となる．

Question 1 では 512 次元，1024 次元ともにベースラインを下回っており，順位ベースの指標であっても十分な結果は得られなかった．コーディネートには一意的な答えがあるわけではなく，トップスにトップスを選びさえすれば回答として大きな問題はない．このため今回の手法での識別は困難であったと考えられる．

一方で Question 2 の場合は，ベースラインを上回っており，トップスの回答としてバッグやシューズなど不適切なものを選ぶことが間違いであることは学習できたことがわかる．正解との距離の計算方法の選び方などを含め，MLP を用いた手法の性能の改善は今後の課題である．

7.1.2 LSTM に基づく提案手法の結果

表 7 に提案手法で学習した時の accuracy と順位指標の計算結果を示す．Question 1 の accuracy は 28.9% となり，MLP でのテスト結果より高い精度が得られた．MLP には 7 つのアイテムの分散表現を連結して 1 つにまとめて入力したが，LSTM では 1 つ 1 つ入力したため，各アイテムの分散表現をより考慮した学習が可能になったのだと考えられる．

また，Question 2 の問題においては，0.510% と MLP と同等の結果が得られた．

順位指標の計算結果に関しても MLP の結果よりわずかに上回っているものの，同等の結果が得られた．

8. 数値実験 3

8.0.1 問題 Question 3 の設定および結果

Question 1 においては MLP および LSTM に基づく提案手法ともにベースラインをわずかに超える結果しか得られなかった．この点について，ネットワーク構成の改良が必要であると考えられるが，まずは問題の難易度を下げることで提案手法の有効性を確認するために問題 Question 3 を設定した．Question 1 では 4 択候補となるアイテムをランダムに選出していたが，Question 3 では true のファッションアイテムの分

表 5: 提案手法による MLP の test accuracy

	512 次元	1024 次元	2048 次元
baseline	0.250	0.250	0.250
Question1 accuracy	0.242	0.250	0.268
Question2 accuracy	0.506	0.511	0.528

表 6: 平均順位指標 （MLP）

	512 次元	1024 次元	2048 次元
baseline	0.438	0.438	0.438
Question1	0.426	0.431	0.447
Question2	0.625	0.630	0.642

散表現から比較的距離が遠い 3 種のアイテムで選択肢を構成した．この問題により，分散表現の距離に基づく学習に意味があるのであれば識別結果の accuracy は向上すると考えられる．

表 8 に Question 3 を LSTM で解いたときの accuracy と順位指標の計算結果を示す．28.9% から 86.5% と精度が大きく上がっているため，分散表現に基づいたファッションアイテムの学習ができていることがわかる．

順位指標の計算結果に関しては，accuracy が 86.5% であるため，不正解だった 13.5% の全データで正解のアイテムが 2 番目の場合，順位指標は 0.932 となり，全て 3 番目の場合，0.899 となる．よって，表 8 より順位指標が 0.930 であることから，ほとんどの正解アイテムは 1 番目か 2 番目であることがわかる．

図 3 に，Question 1 では不正解だったが Question 3 で正解だった問題の例を示す．図 3 に示すように，分散表現の距離が遠いもので選択肢を構成すると，黒のアイテムが選ばれることがほとんどであった．正解のアイテムが黒のアイテムであっても，選択肢も黒であることから，黒のアイテムの分散表現は他の色のアイテムの分散表現から遠い位置にあり，黒同士のアイテムでも形などの色とは違う性質に基づいて潜在空間中に分布していることが考えられる．

9. まとめと今後の課題

本研究ではファッションアイテムの CAE による分散表現に基づくコーディネート理解を目的とし，コーディネート内におけるアイテムの予測をする手法を提案した．結果として，LSTM を用いた提案手法により，本研究のために設定した問題 Question 3 において，高い正解率が得られた．

今後の課題としては，学習器の損失関数に MSE を用いたが，triplet loss のような画像間の類似性を学習する損失関数を使用することや，Polyvore Dataset に含まれる各アイテムのテキスト情報も用いたマルチモーダルな手法を採用すること，分散表現の改良が挙げられる．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) の補助を得て行われたものである．

表 7: 提案手法による LSTM の test accuracy と順位指標

	accuracy	順位指標
baseline	0.250	0.438
Question 1	0.289	0.442
Question 2	0.510	0.644

表 8: Question 3 の test accuracy と順位指標

	accuracy	順位指標
baseline	0.250	0.438
Question 3	0.865	0.930

図 3: Question 1：誤答， Question 3：正答の例

参考文献

[Akiba 19] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, *CoRR*, Vol. abs/1907.10902, (2019)

[Gers 99] Gers, F. A., Schmidhuber, J., and Cummins, F.: Learning to Forget:Continual Prediction with LSTM, *Technical Report IDSIA 01-99* (1999)

[Han 17] Han, X., Wu, Z., Jiang, Y.-G., and Davis, L. S.: Learning Fashion Compatibility with Bidirectional LSTMs, in *ACM Multimedia* (2017)

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation,Vol.9, No.8, pp.1735—1780 (1997)

[Zaremba 15] Zaremba, W.: RECURRENT NEURAL NETWORK REGULARIZATION, *Under review as a conference paper at ICLR 2015* (2015)

[岡谷 15] 岡谷貴之：深層学習 (機械学習プロフェッショナルシリーズ), 講談社 (2015)