CF-015

属性推定を用いた衣服画像ペアの 2 者間関係を表現するキャプションの生成
Generation of Captions Expressing the Bilateral Relationship between Pairs of Clothing Images using Attribute Estimation

阿部 晃平[1)]　横山 想一郎[2)]　山下 倫央[2)]　川村 秀憲[2)]
Kohei Abe　Soichiro Yokoyama　Tomohisa Yamashita　Hidenori Kawamura

1　はじめに

消費者が購入する衣服を選択する際，その基準は異なる消費者の間で多種多様である．また，特定の消費者に限定したとしても，衣服選択の基準を客観的に評価可能な形で定量化することは困難である．こうした理由から，消費者の衣服選択の支援は難しい問題である．消費者は，店頭やオンラインショッピングサイト上の多数の衣服を，自身が持つ特定の基準に基づいて絞り込み，その個々の衣服を評価し，選択する．ここで，絞り込んだ衣服群は同じまたは類似した特徴を持つ衣服の集合となるが，個々の衣服は必ず何かしらの差異を持つ．消費者がこの類似した衣服群から衣服を選択する際の意思決定を支援するための手段として，商品説明文が考えられる．

現状，店舗やオンラインショッピングサイト上で衣服に付与されている商品説明文は，個々の衣服が持つ特性を説明したものである．しかし，上で述べた衣服選択プロセスにおいて，商品説明文は類似した衣服群全体の関係性を考慮したものである必要がある．すなわち，衣服群間の類似性と差異性を適切に表現した内容であるべきだ．このような商品説明文の生成に向けた最初の取り組みとして，本稿では，衣服のペアを対象とし，その 2 者間関係を表現するキャプションの生成手法の提案と評価を行う．

提案手法では，衣服の特徴を表現する属性を画像から推定し，ペア間で比較することで類似性と差異性を明らかにし，キャプション生成に利用した．さらに，アンケート調査を通じて提案手法が本研究の目的に適したキャプションを生成できたかどうかを評価した．その結果，提案手法は，対象となる衣服だけを考慮した単純な画像キャプション生成手法よりも，本研究の目的に適したキャプションを生成できることが確認された．

本稿は以下のように構成される．第 2 章では，本研究の関連研究について概観する．第 3 章では，提案手法について詳細に解説する．第 4 章では，実験に使用した衣服画像データセットとその特性について説明する．第 5 章と第 6 章では，実施した予備実験とメインの実験のプロセスと結果についてそれぞれ述べる．最後に第 7 章で，本稿の結論をまとめ，今後の展望について考察する．

1）　北海道大学 大学院情報科学院
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan
2）　北海道大学 大学院情報科学研究院
Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan

表 1　画像キャプション生成モデルの比較（MS COCO）

モデル	BLEU4	METEOR	画像言語事前学習
NIC[2]	27.7	23.7	無
NICA[5]	25.0	23.9	無
SCST[6]	31.9	25.5	無
ClipCap[7]	33.5	27.5	有
OFA[10]	44.9	32.5	有

2　関連研究

2.1　画像キャプション生成

画像キャプション生成とは，1 枚の入力画像に対して適切な説明テキストを生成するタスクを指す．ここで，ベンチマークデータセット MS COCO[1] における，主要な画像キャプション生成モデルの比較を表 1 に示す．評価指標 BLEU4 と METEOR は，出力されたキャプションと正解のキャプションとの近さを測る自動評価指標であり，値が大きいほどモデルの性能が高いことを示す．Vinyals ら [2] は，CNN[3] と LSTM[4] を組み合わせた深層リカレントアーキテクチャに基づくモデルを提案した．これに続き，Xu ら [5] は，異なる単語を生成する際に，画像中の各領域に注目する注意機構を導入した．さらに，Rennie ら [6] は，強化学習の手法を組み込んだモデルを提案した．近年では，大量の画像とテキストのペアデータを用いて学習した画像言語事前学習モデルがこれまでのモデルよりも高い精度を実現した．Mokady ら [7] は，画像言語事前学習モデル CLIP[8] と言語モデル GPT-2[9] を組み合わせたモデルを提案し，学習時間の短縮と高精度なキャプション生成を実現した．Wang ら [10] は，画像言語事前学習モデルとしては比較的小規模な，2 千万組の画像テキストペアデータを用いた事前学習モデルを提案した．ここで述べたモデルは全て，1 枚の入力画像に対して 1 文のキャプションを生成する．

2.2　画像群に対するキャプション生成

従来の画像キャプション生成に対する応用の一環として，複数の画像に対してキャプションを生成する取り組みがいくつか展開されている．例えば，変化画像キャプション生成という取り組みでは，2 枚の入力画像間で変化した部分を説明するキャプションを 1 文生成する [11, 12]．本研究では，2 枚の入力画像それぞれに

キャプションを生成する．また，従来の画像キャプション生成においては入力画像の特徴的な部分がしばしば無視されるという問題に対処するため，特徴的な画像キャプション生成という取り組みが行われている [13, 14]．これは，1 枚の入力画像をデータベースから取得した類似画像群と比較し，入力画像の特徴的な部分を把握してキャプション生成に活用するというものだ．しかし，この手法では明示的に類似画像を指定できない．本研究では，指定した 2 枚の画像間の関係を表現したキャプションを生成する．

2.3 衣服画像キャプション生成

衣服画像キャプション生成とは，衣服画像を対象とした画像キャプション生成タスクを指す．園田ら [15] は，自ら収集した衣服画像群から入力画像の類似画像を検索し，それによって得られた衣服情報や類似画像の特徴をテンプレートに当てはめてキャプションを生成する手法を提案した．また，Yang ら [16] は，EC サイトにおける商品紹介文の作成支援を目的とし，正確で魅力的なキャプション生成のためのフレームワークを提案した．生成されるキャプションの質を向上させるための指標として，属性レベルの報酬と文レベルの報酬を導入し，最尤推定，属性埋め込み，そして強化学習を用いてモデルの学習を統合した．さらに，約 100 万枚の画像を含む衣服画像キャプション生成のための大規模データセットの構築を行った．

2.4 衣服属性推定

衣服属性推定とは，入力される衣服画像の素材，柄，襟の形状，袖の長さなどの特性を推定するタスクを指す．このタスクは，衣服画像の検索や推薦といった目的で利用されている．Chen ら [19] は，人物が着用する衣服画像とファッション EC サイトの衣服画像とのマッチングを目的とした衣服画像検索タスクに対して，大規模画像データセット ImageNet[20] を学習した CNN[21] と多層パーセプトロン (MLP) を組み合わせたモデルを提案した．同様に，Huang ら [22] も衣服画像検索タスクを視野に入れ，街頭と EC サイトのそれぞれのドメインの画像を処理するために 2 つの CNN をサブネットワークとして持つ深層モデルを提案した．これらのモデルはどちらも衣服領域を表すバウンディングボックスを用いて学習を行う．一方，Liu ら [17] は，袖や襟の位置などのランドマーク情報を学習し，推論時にランドマークの推定を行い，その情報を補助的に用いて属性推定を行うモデルを提案した．ここで，ベンチマークデータセット DeepFashion[17] における衣服属性推定モデルの比較を表 2 に示す．評価指標としては，各画像に対して推定確率上位 k 件の属性を割り当て，正解の属性が何件正しく推定できたかを測定する，Top-k Recall[18] が用いられる．この表から，ランドマーク推定の結果を属性推定に補助的に用いるという手法を導入することで，性能が大きく向上したことが分かる．

表 2　衣服属性推定モデルの比較（DeepFashion）

モデル	Top-3 Recall	Top-5 Recall
WBIT[19]	27.46	35.37
DARN[22]	40.35	50.55
FashionNet[17]	45.52	54.61

3 提案手法

本章では，本研究で提案する衣服画像ペアの 2 者間関係を表現するキャプションの生成手法について述べる．その概要を図 1 に示す．本提案手法は，2 枚の異なる衣服画像 x_A, x_B を入力として受け取り，それぞれに対応するキャプション y_A, y_B をそれぞれに生成する．本提案手法は，以下に示す 3 つのモジュールで構成される．以下の節では，これらのモジュールについて詳しく説明する．

1. 属性推定モジュール
2. 差異の取得モジュール
3. キャプション生成モジュール

3.1 属性推定モジュール

このモジュールでは，衣服属性推定モデル *Pred* を使用して，入力された衣服画像 x_A, x_B の各属性の推定確率 P_A, P_B を出力する．本稿では，FashionNet[17] の DeepFashion 事前学習モデルを衣服属性推定モデル *Pred* として利用する．FashionNet は衣服のランドマークを推定し，その結果を補助的に使用するモデルであり，高い精度が確認されている．推定の対象となる属性は，DeepFashion が持つ 1000 個の属性であり，そのためモジュールの入出力は式 (1), (2) のようになる．ここで，p_A^k, p_B^k は各衣服画像 x_A, x_B に対する k 番目の属性の推定確率を表す．

$$\begin{aligned} Pred(x_A) &= P_A \\ &= (p_A^1,\ p_A^2,\ \cdots,\ p_A^{1000}) \end{aligned} \tag{1}$$

$$\begin{aligned} Pred(x_B) &= P_B \\ &= (p_B^1,\ p_B^2,\ \cdots,\ p_B^{1000}) \end{aligned} \tag{2}$$

3.2 差異取得モジュール

このモジュールでは，属性割り当てアルゴリズム f を使用して，衣服属性推定モデル *Pred* の出力 P_A, P_B を基に，各衣服画像 x_A, x_B に属性集合 $Attr_A, Attr_B$ を割り当てる．ここで，$Attr_A, Attr_B$ を各衣服画像 x_A, x_B の推定属性と呼ぶ．

$$f(P_A) = Attr_A \tag{3}$$

$$f(P_B) = Attr_B \tag{4}$$

次に，推定属性 $Attr_A, Attr_B$ を比較し，両方の衣服画像が持つ属性集合 *CommonAttr* と，一方のみが持つ属性集合 $DiffAttr_A, DiffAttr_B$ を取得する．ここで，*CommonAttr* と $DiffAttr_A, DiffAttr_B$ をそれぞれ，推定共通属性と推定

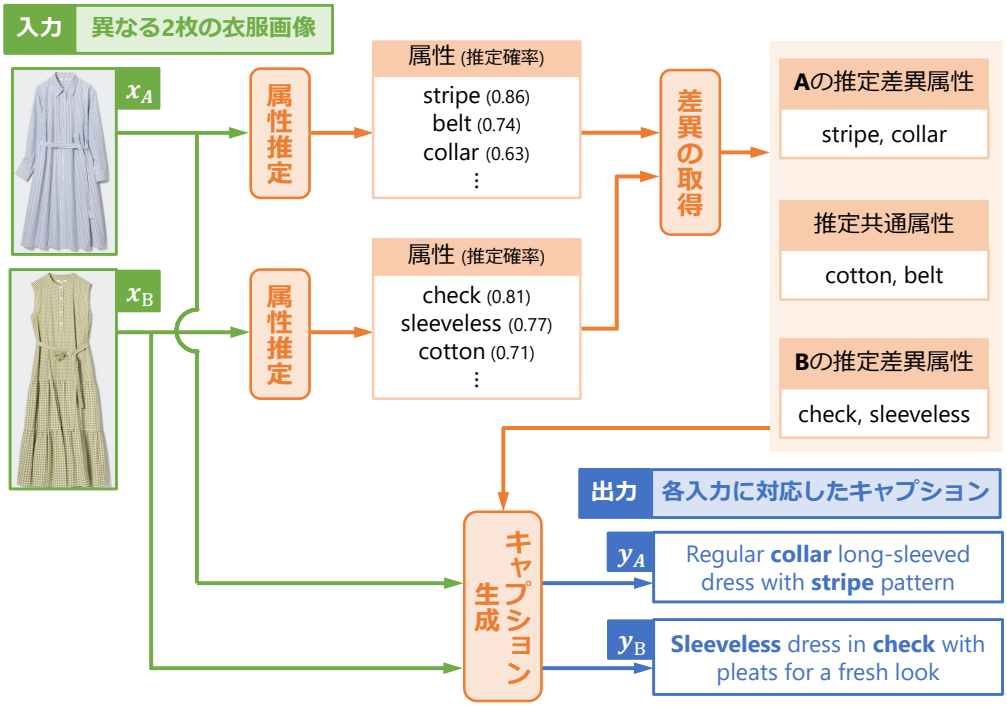

図 1　提案手法の概要

差異属性と呼ぶ．

$$Attr_A \cap Attr_B = CommonAttr \tag{5}$$
$$Attr_A \setminus Attr_B = DiffAttr_A \tag{6}$$
$$Attr_B \setminus Attr_A = DiffAttr_B \tag{7}$$

3.3　キャプション生成モジュール

このモジュールでは，各衣服画像 x_A, x_B に対するキャプション y_A, y_B を 1 つずつ生成する．本稿では，そのための一手法として，画像キャプション生成モデル Cap を使用して各衣服画像 x_A, x_B にキャプション集合 Y_A, Y_B を生成し，その中からキャプション y_A, y_B を選択する方式を取る．キャプション選択には，推定共通属性 $CommonAttr$ と推定差異属性 $DiffAttr_A, DiffAttr_B$ を用いたキャプション選択アルゴリズム g を使用する．

$$Cap(x_A) = Y_A \tag{8}$$
$$Cap(x_B) = Y_B \tag{9}$$
$$g(Y_A) = y_A \tag{10}$$
$$g(Y_B) = y_B \tag{11}$$

ここで，本研究では将来的に日本語でのキャプション生成を目標としていることから，ClipCap[7] を画像キャプション生成モデル Cap として使用する．ClipCap は画像言語事前学習モデル CLIP と言語モデル GPT-2 を組み合わせたモデルであり，その高い精度が確認されている．さらに，CLIP と GPT-2 はどちらも日本語版 [23, 24] が存在し，出力言語の日本語化が実現可能かつ容易である．

表 3　衣服画像データセットの比較

データセット	画像数	属性数	キャプション
FACAD170K[25]	178,849	990	有
DeepFashion[17]	289,222	1,000	無
FashionGen[26]	325,536	-	有
iFashion[27]	1,062,550	228	無
Fashion-MNIST[28]	70,000	-	無

4　データセット

本章では，本研究で使用する衣服画像データセットの選択理由とその詳細について述べる．初めに，主要な衣服画像データセットの比較を表 3 に示す．本研究では，属性推定を用いたキャプション生成を目的としているため，属性とキャプションの両方を含む FACAD170K[25] を利用する．次に，FACAD170K のデータの例を図 2 に示す．各衣服画像は，人物が着用している画像または衣服のみの画像であり，それぞれに英語の 1 文キャプションと英語のメタデータが付与されている．図中の太字の単語は属性を表しており，1 枚の衣服画像には複数の属性が付与されている．

5　予備実験

本章では，3 つの予備実験について述べる．

5.1　属性推定モデルの精度検証

本予備実験では，本稿で対象とするデータの選定を目的として，異なるデータに対する属性推定モデルの精度の検証を行った．

本研究で用いる衣服画像データセット FACAD170K の 990 個の属性と，衣服属性推定モデル FashionNet の事前学習データセット DeepFashion の 1000 個の属性の

キャプション：
A neat **band collar** is beautifully balanced by dropped **shoulder sleeve** in a staple **button** up top **cut** from **stretch** kissed **organic cotton**.

メタデータ：
Front button closure, **Elbow-length** sleeves, 96% **organic cotton**, 4% spandex, Machine **wash**

図 2　FACAD170K のデータ例

表 4　事前学習済み FashionNet の精度

データセット	Top-5 Recall	全属性数
DeepFashion テストデータ	23.55	1000
一致属性を持つ FACAD170K 検証データ	8.76	292
属性当たりの Top-5 Recall が 0 でない一致属性を持つ FACAD170K 検証データ	22.84	26

間で一致する属性は 292 個である．つまり，FashionNet は FACAD170K の 292 個の属性しか推定できない．本稿では，これら 292 個の属性を一致属性と呼ぶ．また，FashionNet で FACAD170K の全ての属性を推定可能にするために再学習を行うには，バウンディングボックスとランドマークのアノテーション付与が必要となる．アノテーションの付与には時間的コストがかかることから，本稿では一致属性を持つデータのみを対象とする．

本予備実験では，一致属性を持つ FACAD170K のデータを事前学習済みの FashionNet に適用した．DeepFashion のテストデータセットと，一致属性を持つ FACAD170K 検証データセットでの FashionNet の属性推定精度を表 4 に示す．FACAD170K での精度は，DeepFashion に比べて大幅に低いことが分かる．属性ごとの推定精度の偏りを調査するため，一致属性ごとに Top-5 Recall を計算した結果，292 個中 266 個の属性の Top-5 Recall が 0 であることを確認した．したがって，Top-5 Recall が 0 でない，つまり属性推定精度が確認できた 26 個の属性を持つ FACAD170K の検証データセットに FashionNet を適用し，その精度を表に示す．また，この 26 個の属性を表 5 に示す．表から，DeepFashion での精度と大きく変わらない精度が確認できた．したがって，本稿ではこの 26 個の属性を対象とし，これらを持つ FACAD170K のデータを使用する．

表 5　本稿で対象とする 26 個の属性

abstract	chiffon	classic	collar	cotton	crochet
denim	faux	fit	flare	floral	graphic
knit	lace	leather	maxi	muscle	pleated
pocket	print	red	skater	sleeve	sleeveless
stripe	wash				

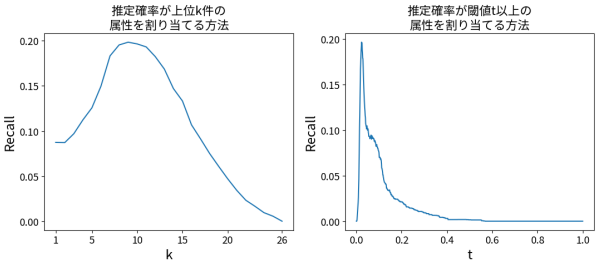

図 3　推定差異属性と正解差異属性との Recall

5.2　属性割り当てアルゴリズムの検証

本予備実験では，本稿で対象とする属性の割り当てアルゴリズム f の選定を目的として，FACAD170K の検証データのペアデータに対して属性を割り当て，推定差異属性 $DiffAttr_A, DiffAttr_B$ と正解差異属性 $DiffAttr_A^{GT}, DiffAttr_B^{GT}$ を比較した．ここで，正解差異属性とは，各衣服画像に正解として付与されている属性集合 $Attr_A^{GT}, Attr_B^{GT}$ を比較した際の，一方のみが持つ属性集合のことを指す．

$$Attr_A^{GT} \setminus Attr_B^{GT} = DiffAttr_A^{GT} \tag{12}$$
$$Attr_B^{GT} \setminus Attr_A^{GT} = DiffAttr_B^{GT} \tag{13}$$

ペアデータとして，正解差異属性と正解共通属性を少なくとも 1 つずつ持つ異なる衣服画像ペアを選定した．本予備実験に使用するペアデータは合計 21,205 ペアである．なお，属性の割り当てによって推定差異属性が現れない場合もあるため，推定差異属性と正解差異属性との Recall を調査した．そして，考慮した割り当て方法は以下の 2 つである．

1. 推定確率が上位 k 件の属性を割り当てる
2. 推定確率が閾値 t 以上の属性を割り当てる

それぞれのパラメータを変化させた際の Recall の値を図 3 に示す．各割り当て方法において，推定確率上位 9 件の属性を割り当てる方法と，推定確率が閾値 0.023 以上の属性を割り当てる方法が最も高い評価を得た．したがって，本稿ではこれらの属性割り当てアルゴリズムを対象とする．

5.3　画像キャプション生成モデルとキャプション選択アルゴリズムの検証

本予備実験では，本稿で対象とする提案手法の設定の選定のため，異なる設定の提案手法の出力キャプションの自動評価を行った．

使用したデータについて説明する．本実験では，対

表 6　ClipCap の学習パラメータ設定

パラメータ	設定値
学習率	2.0×10^{-5}
バッチサイズ	64
エポック数	50

象とする属性を持つ FACAD170K のテストデータから抽出した，22,123 のペアデータを使用した．さらに，画像キャプション生成モデル ClipCap の学習には，FACAD170K の学習データ 177,849 枚を使用した．学習時のパラメータ設定を表 6 に示す．また，各衣服画像に対して 100 文ずつキャプションを生成した．

検証の対象とした設定について説明する．提案手法を適用するには以下の 3 つの設定が必要となる．

1. 属性割り当てアルゴリズム
2. 複数キャプション生成方法
3. キャプション選択アルゴリズム

属性割り当てアルゴリズムとして，予備実験により選定した，推定確率が上位 9 件の属性を割り当てる方法と，推定確率が閾値 0.023 以上の属性を割り当てる方法を用いた．次に，複数キャプションの生成方法として，ビームサイズが 100 のビームサーチと，言語モデルの確率分布に従ったサンプリングを用いた．そして，出力キャプション選択アルゴリズムとして，以下の 3 つの方法を採用した．

1. 推定属性を最も多く含むキャプションを選択
2. 推定差異属性を最も多く含み，推定共通属性を最も少なく含むキャプションを選択
3. 推定差異属性を最も多く含むキャプションを選択

なお，出力の候補となるキャプションが複数ある場合，ビームサーチの場合はスコアが最も高いもの，サンプリングの場合はランダムに 1 文選択した．また，比較のために，FACAD170K を学習した ClipCap を用いてキャプションを生成した．生成方法の設定は Mokady ら [7] を参考に，ビームサイズ 5 のビームサーチとした．

評価指標について説明する．衣服画像ペアの片方だけが正解として持つ属性を正解差異属性，両方が正解として持つ属性を正解共通属性とする．本研究で理想的なキャプションは正解差異属性を多く含み，正解共通属性をできる限り含まないキャプションである．これを考慮して，本実験では以下の 2 つの評価指標を用いた．

1. 出力キャプションに含まれる属性と正解差異属性との F 値
2. 出力キャプションに含まれる属性と正解共通属性との F 値

ここで，指標 1 は大きい方が望ましく，指標 2 は小さい方が望ましい．

表 7　各設定の出力キャプションの評価値

手法	属性割り当て	キャプション生成	キャプション選択		差異 F 値	共通 F 値
			差異	共通		
ClipCap	-	ビームサーチ	-	-	0.224	0.321
提案手法	推定確率上位 9 件	ビームサーチ	多	多	0.285	0.327
				少	0.373	0.226
				-	0.344	0.274
		サンプリング		多	0.292	0.319
				少	**0.385**	**0.202**
				-	0.352	0.255
	推定確率 0.023 以上	ビームサーチ		多	0.285	0.344
				少	0.362	0.298
				-	0.335	0.298
		サンプリング		多	0.294	0.328
				少	0.356	0.227
				-	0.344	0.277

提案手法の各設定における出力キャプションの評価値の平均を表 7 に示す．表 7 から，差異の F 値を見ると，本実験で採用した全ての設定が従来手法を超えていることが分かる．これは，推定差異属性を多く含むキャプションを選択するという設定を適用した結果である．また，属性割り当てアルゴリズムとして推定確率上位 9 件の属性を割り当てる方法，そして複数キャプション生成方法としてサンプリング，出力キャプション選択アルゴリズムとして推定差異属性を最も多く含む一方で，推定共通属性を最も少なく含むキャプションを選択する方法を採用した提案手法が，本実験の 12 通りの設定の中で最も高い評価値を得た．以上の結果に基づき，本稿ではこの設定を用いた提案手法を対象とする．

6　実験

6.1　実験目的

本実験では，提案手法が衣服画像ペアの 2 者間関係を適切に表現するキャプション生成手法であるかどうかの検証と，類似衣服画像ペアと非類似衣服画像ペアにおける提案手法の効果の比較を目的とする．

6.2　実験方法

本実験では，提案手法と従来手法 ClipCap の出力キャプションについてアンケート調査を実施した．アンケートの対象者は，20 代の男女 15 人である．アンケートでは，衣服画像ペアを 5 組提示した．これらは全て，属性推定で正解となる衣服画像である．5 組のペアのうち，類似衣服画像ペアが 3 組，非類似衣服画像ペアが 2 組である．本稿では，これらをそれぞれ類似ペア，非類似ペアと呼ぶ．これらのペアは，各衣服画像が正解として持つ属性の類似度と目視による確認を基に選択した．また，キャプションは予備実験で最も高い評価を得た設定を使用した提案手法の出力キャプションと，従来手法の出力キャプションの 2 組を提示した．したがって，衣服画像ペアとキャプションのセットは合計 10 組を提示した．出力キャプションはどちらの手法も英語のため，英文読解を補助するために日本語訳も表示した．実際に表示した衣服画像ペアとキャプションのセットを図 4, 5 に

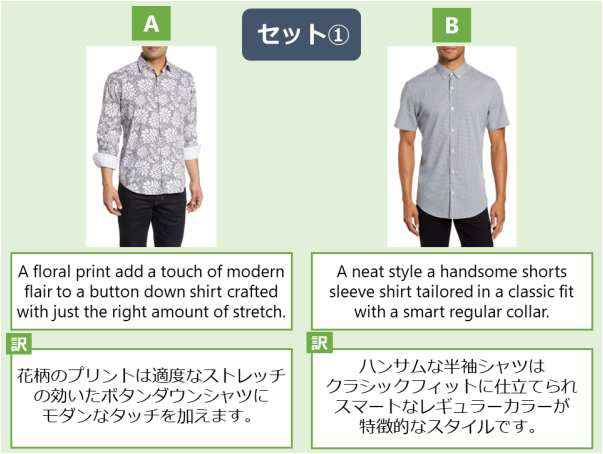

図 4　アンケートに提示した画像とキャプション（提案手法）

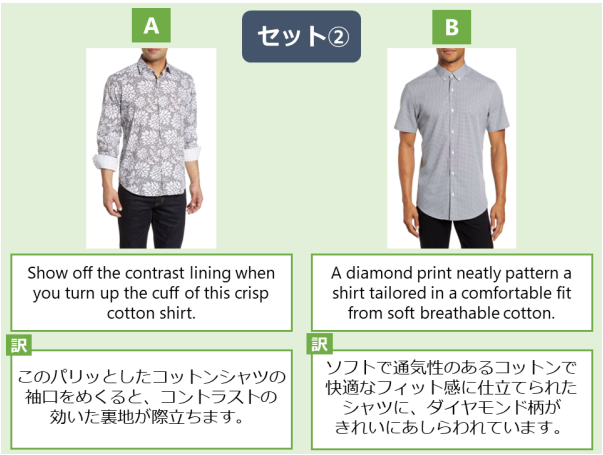

図 5　アンケートに提示した画像とキャプション（従来手法）

示す．そして，調査結果の検定には Z 検定を使用し，有意水準は 5%に設定した．

アンケートで設定した質問とその選択肢を表 8 に示す．衣服画像ペア間で差異のある特徴と共通する特徴がキャプションに記述されているかどうかを調査するために，問 1 と問 2 を設定した．さらに，キャプションが本研究の目的に適しているかどうかを調査するために，問 3 を設定した．

6.3　実験結果

各質問に対する回答の割合を図 6〜8 に示す．また，類似ペアと非類似ペアでの問 3 の回答の割合を図に示す．

図 6 から分かるように，差異のある特徴に関する記述が 1 つ以上あると回答された割合では，提案手法が 92.3%で，従来手法が 67.6%であり，明らかな差が見られる．この差は検定により有意であることを確認した．したがって，提案手法は従来手法に比べて，差異のある特徴を記述したキャプションをより多く生成できたといえる．

図 7 からは，共通する特徴に関する記述が無いと回答された割合で，提案手法でが 63.0%，従来手法が 31.5%

表 8　アンケートの質問と選択肢

	質問	選択肢
問 1	ペア間で差異のある特徴に関する記述があるか．	複数ある 1 つだけある 無い
問 2	ペア間で共通する特徴に関する記述があるか．	複数ある 1 つだけある 無い
問 3	ペア間の差異を強調したキャプションとして適切か．	適切 どちらかといえば適切 どちらかといえば不適切 不適切

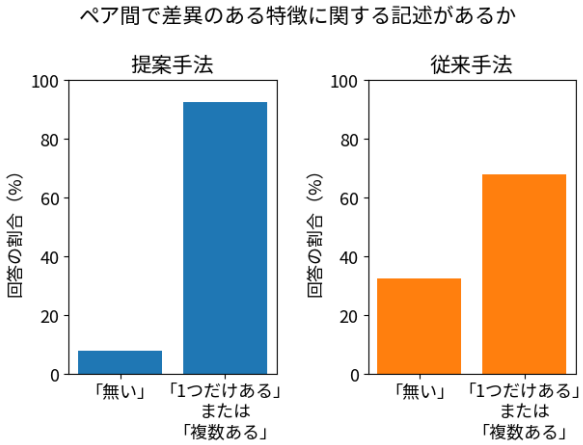

図 6　問 1 に対する回答の割合

という差が見られる．これも検定により有意な差であることを確認した．したがって，提案手法は従来手法に比べて，共通する特徴を記述しないキャプションをより多く生成できたといえる．

図 8 によれば，差異を強調したキャプションとして「適切」または「どちらかといえば適切」と回答された割合では，提案手法が 89.2%で，従来手法が 47.3%と大きな差がある．これも検定により有意な差であることを確認した．したがって，提案手法は従来手法に比べて，差異を強調したキャプションとして適切と感じられるキャプションをより多く生成できたといえる．

図 9, 10 からは，差異を強調したキャプションとして「適切」または「どちらかといえば適切」と回答された割合について，類似ペアでの提案手法が 92.3%，従来手法が 43.5%と大きな差が見られるのに対して，非類似ペアでの提案手法が 84.6%，従来手法が 69.2%とそれほど大きな差は見られない．これから，提案手法は非類似ペアよりも類似ペアでの有効性が高いといえる．これは，非類似ペアが類似ペアよりも差異のある特徴が多く，単純な画像キャプション生成手法である従来手法でも十分に差異のある特徴を記述できたためと考えられる．

7　おわりに

本研究では，消費者が類似した衣服群から購入する衣服を選択する際の意思決定支援に向けて，衣服群の関係を考慮した商品説明文の自動生成を提案した．本稿で

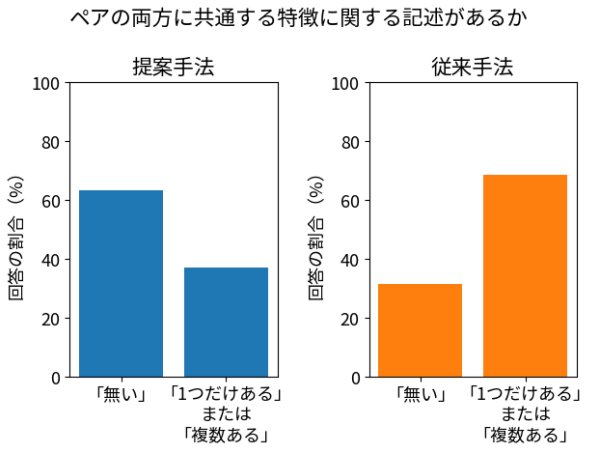

図 7　問 2 に対する回答の割合

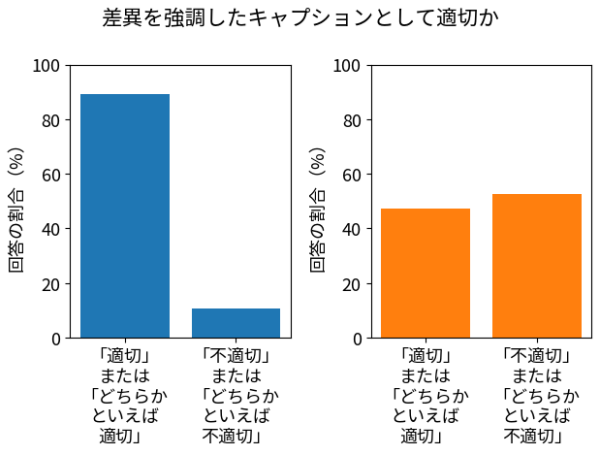

図 8　問 3 に対する回答の割合

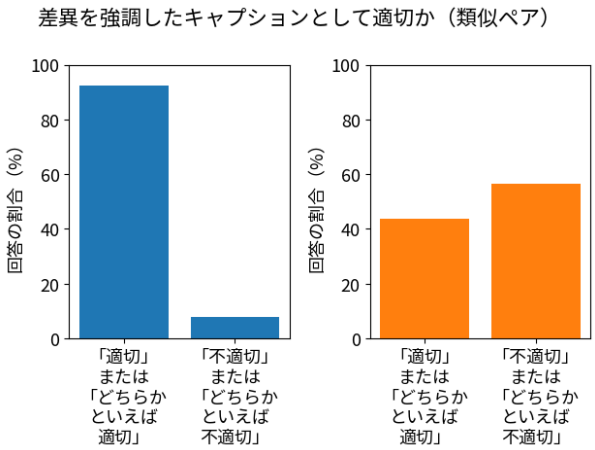

図 9　問 3 に対する回答の割合（類似ペア）

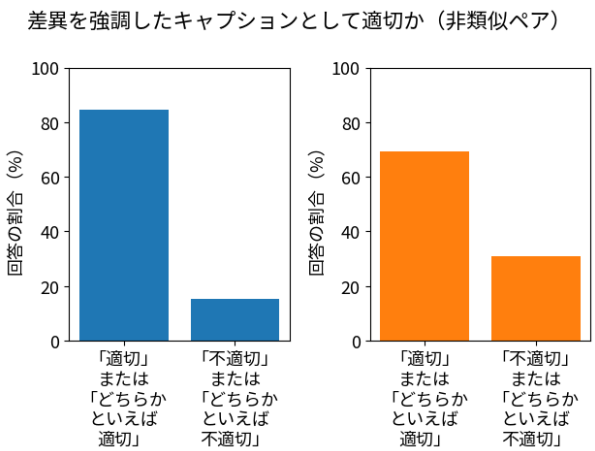

図 10　問 3 に対する回答の割合（非類似ペア）

は，その最初の取り組みとして，衣服画像ペアの 2 者間関係を表現するキャプション生成手法の提案と評価を行った．

提案手法は，入力となる衣服画像からその属性を推定し，これらをペア間で比較することで，共通点と差異を特定した．その情報を利用し，各画像に対して独立に生成された複数のキャプションの中から，最終的なキャプションを選択した．アンケート調査による評価では，提案手法が本研究の目的により適したキャプションを生成できることが示された．また，提案手法は，非類似ペアよりも類似ペアに対して高い効果を示すこと確認した．

本稿で対象とした属性は限られており，すべての衣服特徴をカバーしているわけではない．また，3 枚以上の衣服画像を入力とする場合には，取り扱える属性の数が制約となる可能性がある．そのため，今後の研究では，扱える属性の数を増やすことや，より多くの衣服画像を処理できるように手法の拡張に取り組むことが必要とされる．

参考文献

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick, “Microsoft coco: Common objects in context”, European conference on computer vision, pages 740–755 (2014).

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, “Show and tell: A neural image caption generator”, Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164 (2015).

[3] Sergey Ioffe, Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, International conference on machine learning, pages 448–456 (2015).

[4] Sepp Hochreiter, Jürgen Schmidhuber, “Long short-term memory”, Neural computation, 9(8):1735–1780 (1997).

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio, “Show, attend and tell: Neural image caption generation with visual attention”, International conference on machine learning, pages 2048–2057 (2015).

[6] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, Vaibhava Goel, “Self-critical sequence training for image captioning”, Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7008–7024 (2017).

[7] Ron Mokady, Amir Hertz, Amit H Bermano, “Clipcap: Clip prefix for image captioning”, arXiv:2111.09734 (2021).

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya

Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, “Learning transferable visual models from natural language supervision”, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763 (2021).

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, “Language models are unsupervised multitask learners”, OpenAI blog, 1(8):9 (2019).

[10] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, Hongxia Yang, “Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework”, International Conference on Machine Learning, pages 23318–23340 (2022).

[11] Harsh Jhamtani, Taylor Berg-Kirkpatrick, “Learning to describe differences between pairs of similar images”, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4024–4034 (2018).

[12] Dong Huk Park, Trevor Darrell, Anna Rohrbach, “Robust change captioning”, Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4624–4633 (2019).

[13] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, Antoni B Chan, “Group-based distinctive image captioning with memory attention”, Proceedings of the 29th ACM International Conference on Multimedia, pages 5020–5028 (2021).

[14] Yangjun Mao, Long Chen, Zhihong Jiang, Dong Zhang, Zhimeng Zhang, Jian Shao, Jun Xiao, “Rethinking the reference-based distinctive image captioning”, Proceedings of the 30th ACM International Conference on Multimedia, pages 4374–4384 (2022).

[15] 園田亜斗夢, 新名玄, “アパレル EC サイトにおける説明文自動生成”, 経営情報学会 全国研究発表大会要旨集 2018 年秋季全国研究発表大会, pages 125–127 (2018).

[16] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, Xin Wang, “Fashion captioning: Towards generating accurate descriptions with semantic rewards”, Computer Vision – ECCV 2020, pages 1–17 (2020).

[17] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, Xiaoou Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations”, Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1096–1104 (2016).

[18] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, Sergey Ioffe, “Deep convolutional ranking for multilabel image annotation”, arXiv:1312.4894 (2013).

[19] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, Tamara L. Berg, “Where to buy it: Matching street clothing photos in online shops”, Proceedings of the IEEE International Conference on Computer Vision (2015).

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database”, IEEE Conference on Computer Vision and Pattern Recognition, pages 248-255 (2009).

[21] Karen Simonyan, Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition”, arXiv:1409.1556 (2014).

[22] Junshi Huang, Rogerio S Feris, Qiang Chen, Shuicheng Yan, “Cross-domain image retrieval with a dual attribute-aware ranking network”, Proceedings of the IEEE international conference on computer vision, pages 1062–1070 (2015).

[23] rinna 株式会社, “japanese-clip-vit-b-16”, `https://huggingface.co/rinna/japanese-clip-vit-b-16` (accessd 2022).

[24] rinna 株式会社, “japanese-gpt2-medium”, `https://huggingface.co/rinna/japanese-gpt2-medium` (accessd 2022).

[25] Chen Cai, Kim-Hui Yap, Suchen Wang, “Attribute conditioned fashion image captioning”, IEEE International Conference on Image Processing, pages 1921–1925 (2022).

[26] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, Chris Pal, “Fashion-gen: The generative fashion dataset and challenge”, arXiv:1806.08317 (2018).

[27] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R Scott, Serge Belongie, “The imaterialist fashion attribute dataset”, Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019).

[28] Han Xiao, Kashif Rasul, Roland Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms”, arXiv:1708.07747 (2017).