The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

MAGMA を利用したファッションコーディネートのためのコメント生成

Comments generation using MAGMA for fashion outfit

青葉 紗矢香 *1 Sayaka Aoba　佐藤 勇元 *1 Yugen Sato　益川 良藏 *1 Ryozo Masukawa　佐藤 真 *1 Makoto Sato　櫨 翔佑 *1 Shosuke Haji　松井 太我 *2 Taiga Matsui
石川 桂太 *2 Keita Ishikawa　高木 友博 *1 Tomohiro Takagi

*1 明治大学理工学部情報科学科
Department of Information Science, School of Science and Technology, Meiji University

*2株式会社エアークローゼット
airCloset, Inc.

In this study we build a model to generate comments when recommending two coordinated fashion items in personal styling. In coordinating recommendations, advice is also provided on the compatibility of these items and their combination with accessories and other items. We uses MAGMA, a method that supports multimodal input for language models by using adapters, to generate coordinating comment sentences with the combined image of two clothes and prompts as input. Quantitative and qualitative comparisons with manually generated comments and conventional methods based on machine learning were conducted to confirm that the proposed model is fully capable.

1. はじめに

近年，画像と言語の両方を特徴量とする機械学習モデルが注目されている．それらの手法は複数のモーダルを扱うことにより，単一のモーダルのみを扱う場合よりも優れた成果を挙げている．

パーソナルスタイリングにおいて，顧客が配送されたファッションアイテムの選択理由を容易に理解できるよう，スタイリストは顧客情報，トレンドなどの様々な要因からコーディネートの説明文を作成する．ファッションコーディネートの理解を容易にするためにはキャッチーなフレーズを用いることが効果的であり，ファッションコーディネートを端的かつ正しく説明するためには“組み合わせるアイテム個々の特徴”と“アイテムの相性”の両方を捉えることが必要がある．

本研究では画像と言語の両方を利用し，パーソナルスタイリングにおいて 2 つのファッションアイテムのコーディネートを説明する際のコメントに含める，コーディネイトを端的に表したフレーズを生成するモデルを構築する．具体的には，言語モデルをマルチモーダルな入力に対応させた手法である MAGMA を利用し，2 枚の服画像を結合した画像と特徴的なフレーズを含むプロンプトを入力として，コーディネートに関する説明をするフレーズを生成する.MAGMA は Adapter を採用しており，Adapter で特定のドメイン知識を得ることで学習パラメータの数を減らし学習コストを削減する．

人手，ラベル分類モデル，提案手法それぞれが生成したフレーズを用いて定量・定性比較を行ない，提案モデルが十分な能力を持つことを示す．

2. 関連研究

商品説明の自動生成：

商品説明には購買意欲を高めるためのキャッチーなフレーズが用いられる．[Vitobha Munigala 18] では，入力された短い商品仕様からキーワードを抽出してタグ付けし，そのタグとファンション用語リスト内の単語との単語間類似度を用いて更にタグを拡張し，それらのタグに類似したフレーズと言語モデルを利用して説得力のある文章の生成を行っている．

3. 提案システム

3.1 システム構成

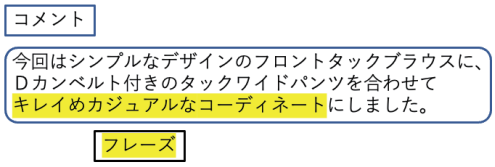

図 1: コメントに含まれるフレーズの具体例

本研究では，MAGMA を利用して 2 つのファッションアイテムのコーディネートを端的に表すフレーズを生成するシステムを構成する．ここで言うフレーズとは，パーソナルスタイリングで作成される，顧客に配送されるコメント内に含まれるものであり，図 1 に示した“キレイめカジュアルなコーディネート”が例である．提案システムの構成を図 2 に示す．提案システムでは，トップス画像とボトムス画像を結合した 1 枚の画像と，2 つのアイテム名を含む質問形式のプロンプトを入力とする．この入力される 2 枚の画像は配送される 2 アイテムに対応する．質問形式のプロンプトに対する回答として，フレーズを生成する．

本研究では 1 つの入力に対していくつかのフレーズを生成し，その中から最適な 5 個のフレーズを選択することでフレーズ候補群を作成した．

連絡先: 青葉紗矢香，明治大学理工学部情報科学科，〒 214-8571 神奈川県川崎市多摩区東三田 1-1-1, aoba_@cs.meiji.ac.jp

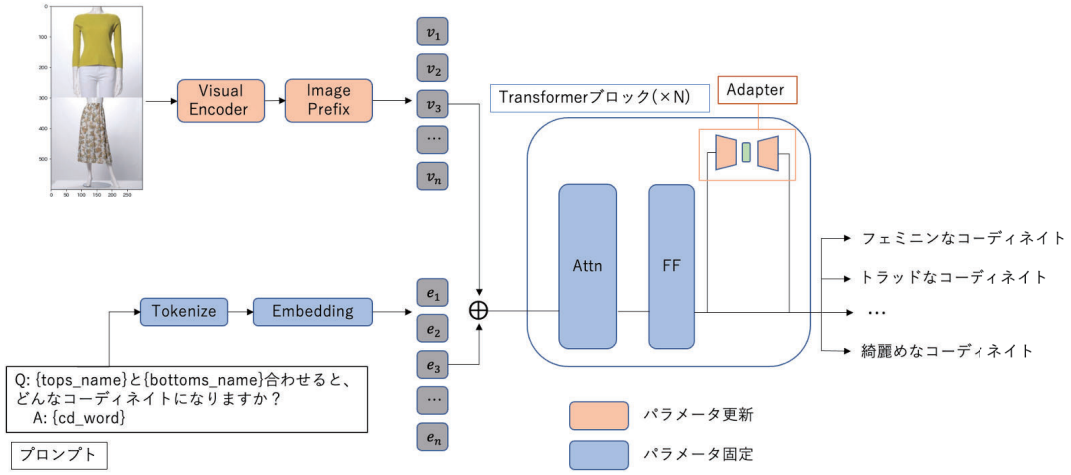

図 2: MAGMA を利用した提案システムの構造

3.2 プロンプト

前述したように, 提案システムでは画像の他にプロンプトを入力として扱う. 本研究では学習時のプロンプトを“Q: {トップスの名前} と {ボトムスの名前} 合わせると、どんなコーディネートになりますか？ A: {コーディネートを表すフレーズ} ”の形式で統一する. トップスとボトムスの特徴を別々に捉えることを目的に, プロンプトにアイテム名を含める. 生成時には, 学習時のプロンプトから“{コーディネートを表すフレーズ} ”を取り除いた“A: ”までのプロンプトを入力に用い, 質問形式のプロンプトの答えとして 2 つのアイテムのコーディネートを表すフレーズを生成する.

3.3 MAGMA

提案システムで利用する Multimodal Augmentation of Generative Models through Adaptebased Finetuning(MAGMA)[Constantin Eichenberg 22] を利用して, 画像と言語の任意の組から学習を行うモデルである.MAGMA は学習中に言語モデルの重みを更新しないため, 巨大言語モデルが得た知識を転用することが可能である. また, Adapter ベースのファインチューニングを行うことで学習パラメータの数を減らし学習コストを削減している. MAGMA は幅広いベンチマークにおいて最新の Vision and Language モデルと競合する性能を持ち, 特に外部知識を必要とするタスクにおいて優れた性能を発揮する.

MAGMA は主に 4 つの構成要素に分けることができ, 本節ではこの 4 つのモジュールを簡潔に説明する.

3.3.1 Visual Encoder

Visual Encoder は入力画像の意味情報を抽出するためのモジュールである. 画像と文を比較することにより事前学習された CLIP[Alec Radford 21] の Visual Encoder を利用している.Visual Encoder は入力画像を処理して特徴ベクトルの系列を出力し, Image Prefix の入力とする.

3.3.2 Image Prefix

Image Prefix は, Visual Encoder の出力を Language Model へ入力するために追加で処理を行うモジュールである.CLIP エンコーダーの場合, 出力が $N \times N$ グリッドのため, これを N^2 個のベクトル列に平坦化し, Language Model の隠れ次元数である d_h 次元に線形変換を行う.

3.3.3 Language Model

Language Model は Tokenize や Embedding を行うモジュールと Transformer ブロックから成り, 入力された特徴を元に文生成を行う.Language Model ではまず入力文がトークンのシーケンスに変換される. 更に, 単語埋め込み層を利用してそれぞれのトークンがベクトルに変換され, Transformer ブロックの Decoder モジュールに入力される. しかし元の MAGMA は英語で事前学習されているため日本語の生成には不向きである. そこで本研究では, GPT-J を日本語 GPT-2[rinna Co., Ltd. 22] に置き換え日本語文の生成を行う.

3.3.4 Adapter

Adapter は Transformer ブロックの要素間に配置されるモジュールである.Language Model を学習する代わりに Adapter を中心とした学習を行うことで, 学習に必要なパラメータ数を減らしモデルをより効率的に最適化する.

4. 学習

学習に用いるデータは, トップスとボトムスのアイテム画像 1 枚ずつを上下に結合したコーディネート画像 1 枚と, 3.2 節で示したプロンプトの組である.

MAGMA のファインチューニングを行う際, 図 2 にある通り Tokenize や Embedding を行うモジュールと Transformer ブロックからなる Language Model の重みは固定され, Visual Encoder, Image Prefix, Adapter の重みのみは更新する.Language Model は事前学習済み日本語 GPT-2 により初期化され, Visual Encoder は事前学習済み CLIP の重みで初期化される. また, Image Prefix と Adapter は 0 から学習される.

以下, モジュールの学習可能なパラメータを θ で表す. 画像とキャプションの組 (x, y) が与えられた時, 画像は式 (1), キャプションは式 (2) に従って埋め込み表現を獲得する.

$$v_{1,\theta}, \cdots, v_{n,\theta} = V_\theta^p \circ V_\theta^e(x) \tag{1}$$

$$e_1, \cdots, e_n = E(t_1), \cdots, E(t_m) \tag{2}$$

式 (1),(2) 内の V^p は Image Prefix の出力ベクトル, V^e は Visual Encoder の出力ベクトル, E は単語埋め込み層, t_k はキャプション y をトークン化したものである. また, 画像列の

表 1: フレーズの生成例

プロンプト	生成フレーズ
Q: シンプルなデザインのフロントタックブラウスとDカンベルト付きのタックワイドパンツ合わせると、どんなコーディネートになりますか？ A:	きれいめカジュアルなコーディネート
	爽やかなスカートコーディネート
	気負わない雰囲気のコーディネート
	スタイリッシュな大人フェミニンコーディネート
	きれいめなスカートコーディネート

図 3: 入力アイテムペア

長さ n は固定だが, キャプションの長さ m はパディングを用いるため可変である. この画像埋め込みとキャプション埋め込みを結合し Transformer ブロックの入力とする.

損失関数は式 (3) を用いる.

$$L_\theta(x, y) = -\sum_{i=1}^{m} l_\theta(v_{1,\theta}, \cdots, v_{n,\theta}, e_1, \cdots, e_n) \qquad (3)$$

式 (3) において l_θ は, Transformer ブロックの最後の埋め込みを, 次のトークンを予測する際に利用するトークン空間上に写像したベクトル H と, Adapter を追加した新たな Decoder モジュールのベクトルである $\tilde{T}$ より算出される.

5. 実験

本章では提案システムを定性・定量的に検証する. 実験は, パーソナルスタイリングにおいてトップスとボトムスの 2 アイテムが顧客に送られる場合に, 配送される 2 アイテムに付与されるコメントに含めるコーディネートを表すフレーズを生成する, という問題設定で行う. 実験時には画像とプロンプトの組を入力しフレーズを生成する.

5.1 データセット

本研究で用いるデータセットは, スタイリストが実際に顧客に送った過去の配送データから作成した約 100 万件である. この配送データから実験に必要なアイテム情報を抜き出し, 適切に前処理を行った上で利用した. 具体的には, コーディネートを端的に表したフレーズを配送に付けられるコメントから, アイテム名をあらかじめアイテムに付けられているコメントから抜き出し利用した. ここで言うアイテム名とは, “ シンプルなデザインのフロントタックブラウス”や“ Dカンベルト付きのタックワイドパンツ”など, アイテムのディティールにまで言及した詳細な名称のことである.

また提案システムではトレンドなどの時期により変化する要因を扱わないため, それに関連するフレーズは除去した.

5.2 比較手法

提案システムとの比較手法として, 本実験では学習済みの VisionTransformer[Alexey Dosovitskiy 21] と EfficientNet[Mingxing Tan 19] を組み合わせた, トップス画像 1 枚とボトムス画像 1 枚の計 2 枚の画像のみを入力とした, 188 ラベル分類モデルを用いる. 図 4 にラベルの作成方法を示す. ラベルは 2 アイテムの配送に付けられるコメントデータから作成され, 具体的には “ 綺麗”や“ フェミニン”などコーディネートを表す単語となっている. 実際にコメントに含めるフレーズには, ラベルの元となった, コメントデータから引用した詳細なフレーズである “ キレイめカジュアルなコーディネート”や“ レディライクなコーディネートト”を利用する.

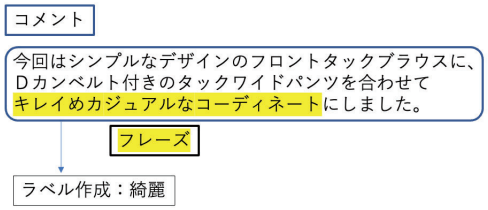

図 4: 多ラベル分類で利用するラベルとフレーズ

5.3 実験結果

図 3 の入力に対するフレーズの生成例を表 1 に示した. これより, 同様の形式で意味の方向性が異なるフレーズを生成できたことが確認できる.

6. 評価

6.1 定量評価

提案モデルの評価をするにあたって, 要約タスクなどの評価に利用される ROUGE-1, BERT を利用した BERTScore[Jacob Devlin 19], 翻訳タスクなどの評価に利用される BLEU を用いた. 結果を表 2 に示す. 太字は各評価指標で最も精度が高いことを表す.

表 2: 定量評価

モデル	ROUGE-1	BERTscore	BLEU
分類モデル	0	0.654	0.265
提案システム	**0.192**	**0.786**	**1.037**

表 3: 定性評価

モデル	日本語評価	ファッション評価	最良選択割合 (郡)	最良選択割合 (個別)
分類モデル	**2.422**	1.256	0.392	0.236
提案システム	2.404	**1.991**	**0.741**	**0.719**

6.2 定性評価

特徴的な用語が頻出するファションの日本語文を定量評価のみで測るのは難しい. そこで本研究では, 2 つのモデルが出力したフレーズをそれぞれ 5 つずつ, 計 10 個の生成フレーズに対してスタイリストによる人手評価を実施した. 具体的には, 各手法が生成したフレーズを日本語的に正しいか, ファッション面で優れているかの 2 つの側面で, 0-3 点の数値で評価した.2 種類の評価指標でスコアリングすることによって日本語の言い回しや文法的な正しさに対する評価を行うと共に, 2 アイテムのコーディネートに対して生成されたフレーズのファッション的な正しさも評価する。

加えて, 1 つの配送に対して各手法が生成した 5 つのフレーズを含む候補群のうちどれが最も望ましい候補群であるかを選択する評価を実施した. これにより, どの手法が最も平均的に優れたフレーズを生成できるかを評価した. また群に含まれる 10 の個々のフレーズのうち最も望ましい生成がどれかを選択する評価も実施した. これにより, 最も優れた 1 つの生成フレーズと最も優れた生成群が一致するか確認した.

全 15 アイテムを 3 名のスタイリストによって評価した結果を表 3 に示す. 太字は各評価指標で最も精度が高いことを表す.

6.3 考察

表 2 より, 定量評価においては提案システムがすべての指標で最も良い精度となった. 共通語の数に注目する ROUGE や BLUE, 単語の意味に注目する BERT がどちらも最も良い結果であることから, 提案システムが最も正解のフレーズを再現できていると言える.

表 3 より, 定性評価においてはファッション評価と最良選択の, 群, 個別の両方において提案システムが分類モデルの精度を上回った. よって, 提案システムは分類モデルと比較した場合に, 実運用により適した候補, 候補群を生成可能であるとわかる. 一方日本語評価においては分類モデルが最も良い精度となった. 分類モデルにおいては人手で作成されたフレーズを選択して利用するため, 日本語の正しさは担保されている. よって生成を行う提案システムと比較した場合に言い回しや単語の間違いが減り, 提案システムの精度を上回ったと考えられる.

これらの結果より, 提案システムではコメントデータから引用されたフレーズと同程度の精度の日本語表現で, よりコーディネートに適したフレーズを生成可能であると言える.

7. おわりに

本研究では, Adapter ベースの手法である MAGMA をファインチューニングすることにより, 2 枚のアイテム画像を結合した 1 枚の画像とアイテム名を含むプロンプトからコーディネートを端的に表すフレーズを生成することに成功した. また実験の結果, 提案システムは定性的に見て十分な性能を持つことがわかった. 定量評価が難しいファッション分野でより正確な日本語文を生成するためには, ユーザー情報やトレンド情報なども扱うことが必要である. しかし提案システムで利用可能なプロンプトに言語で更に情報を追加しても, 特徴抽出の過程で各情報を個別に保持し続けるのは難しいと考えられるため, より多くの情報を個別に扱うための改善が必要である.

参考文献

[Constantin Eichenberg 22]Letitia Parcalabescu and Anette Frank: Heidelberg University: MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning(2022)

[Alec Radford 21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever: Equal contribution, OpenAI, : Learning Transferable Visual Models From Natural Language Supervision(2021)

[rinna Co., Ltd. 22]rinna Co., Ltd. : https://huggingface.co/rinna/japanese-gpt-1b(2022)
[Vitobha Munigala 18]Vitobha Munigala, Abhijit Mishra, Srikanth G. Tamilselvam, Shreya Khare, Riddhiman Dasgupta, Anush Sankaran: PersuAIDE ! An Adaptive Persuasive Text Generation System for Fashion Domain(2018)

[Alexey Dosovitskiy 21]Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale(2021)

[Mingxing Tan 19]Mingxing Tan, Quoc V. Le: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks(2019)

[Jacob Devlin 19]Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi : Department of Computer Science and Cornell Tech, Cornell University, ASAPP Inc : BERTScore: Evaluating Text Generation with BERT(2019)