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

ファッションアイテム推薦のための大規模言語モデルを用いたコメント生成

Comment generation using a large-scale language model for fashion item recommendation

佐藤勇元 *1 Yugen Sato　青葉紗矢香 *1 Sayaka Aoba　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 personal styling,the stylist selects fashion items based on the client's characteristics, purpose of use, season, and various other factors. The stylist then carefully comments on the reasons for the selection and sends it to the customer along with the recommended fashion item. In response to this, we use MAGMA, a method that supports multimodal input of language models by means of adapter-based fine tuning, to construct a model that generates comment suggestions based on the combination of item images and prompts. We conducted quantitative and qualitative evaluations of the proposed model and confirmed that the model using MAGMA is superior to the conventional method.

1. はじめに

近年，大規模モデルに対する注目は高まっており，それに関する研究も行われている．また，画像と言語の双方を扱うマルチモーダル処理についても注目を集めている．本研究ではマルチモーダルな大規模モデルをファッションドメインへ適応させることを目的とする．

パーソナルスタイリングにおいて，スタイリストは顧客の特徴，利用目的，季節などの様々な状態を考慮してファッションアイテムを選択するが，その際その選択理由について注意深くコメントし，推薦するファッションアイテムとともに顧客に送る．このコメント生成タスクに対し我々は，Adapters ベースのファインチューニングによって言語モデルをマルチモーダル入力に対応させた手法である MAGMA を利用し，アイテム画像の組み合わせとプロンプトから，コメント案を生成するシステムを構築する．大規模モデルのファインチューニングは通常コストが大きいが，Adapters(3.2.4 に後述) を用いることでコストを抑え，かつ高い生成精度を実現できる．また，チューニングの際は画像とテキストプロンプトのペアで学習をするため，プロンプトの内容によって生成を制御することが可能である．提案システムではプロンプトで注目点のカテゴリを指定することでそれに沿った内容のコメントを生成する．

提案システムに対して定量・定性評価を行った結果，従来手法よりも優れていること，特定の条件下にて人手と比較しても遜色ない能力を持つことを確認した．

2. 関連研究

ファッション表現を獲得するための機械学習:
ファッションドメインは定量的に評価するのが難しい．例えばコーディネートの良し悪しやファッションの方向性 (テイスト) を判定する場合など，指標のない中でファッション特有の曖昧な表現が多く使用される．[Ryotaro Shimizu 23] では全身コーディネート画像と画像に付与された複数のタグ情報を同一の空間に写像し，この空間における画像とタグの座標（=埋め込み表現）を活用することで，曖昧なファッション表現を理解してユーザの問い合わせに正確に応答するシステムを構築している．

連絡先: 佐藤勇元，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,

3. 提案システム

3.1 システム概要

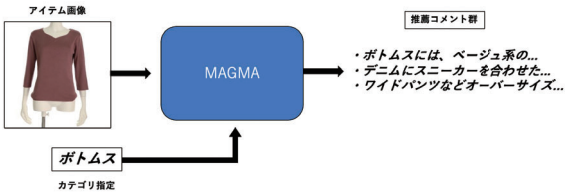

図 1: システム全体像

システムの全体像と動作例を図 1 に示す．
ここで,「アイテム画像」は顧客に送るファッションアイテムの画像,「ボトムス」は画像に示されたアイテムにどのようなボトムスが似合うかを考える状況を示しており,「 推薦コメント」はその状況でシステムが生成したコメントを示す．ボトムス以外に、トップス，シューズ，アクセサリーの 3 種類を合わせ，全部で 4 種類のカテゴリを組み合わせ対象として想定する．コメント候補は 30 文生成し，3 クラスタにクラスタリングした上で，各クラスターから 1 文ずつ抽出することでコメント候補の最終出力とする．

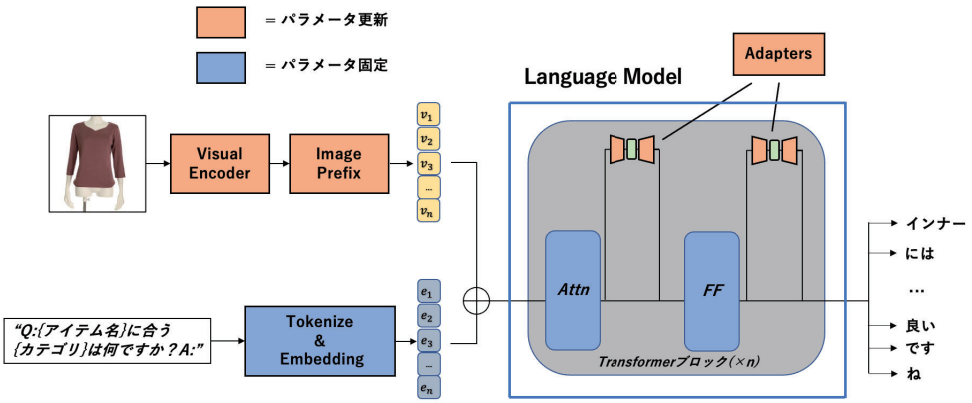

図 2: MAGMA アーキテクチャ

3.2 MAGMA

Multimodal Augmentation of Generative Models through Adapter-based Finetuning(MAGMA)[Constantin Eichenberg 22] は，言語生成モデルをマルチモーダル入力に対応させたモデルであり，オープンエンドな生成タスクや OKVQA ベンチマークで高い精度を実現している．
MAGMA のアーキテクチャは主に 4 つのモジュールで構成されている (図 2)．本研究では簡潔にその構成要素を紹介する．

3.2.1 Visual Encoder

Visual Encoder は入力画像を処理するモジュールである．ここでは画像特徴とテキスト特徴を同一空間上で近づけるよう事前学習されたモデルである CLIP[Alec Radford 21] のエンコーダ部分を用いることで，画像に関する意味情報を抽出する．

3.2.2 Image Prefix

Image Prefix は，Visual Encoder の出力を後の Language Model(3.2.3 に後述) の入力に繋ぐ役割を担っており，画像の特徴を一連の埋め込みベクトルにマッピングする．Visual Encoder 出力は $N \times N$ グリッドのため，N^2 のベクトルに平行化し，LM の隠れ次元 d_h に線形変換を施すことで，画像特徴量の出力とする．

3.2.3 Language Model

言語モデル部分は GPT([Alec Radford 18]) と同様に，Transformer を利用した自己回帰型モデルを使用する．そのため，学習済み GPT の重みを転用することが可能である．[Constantin Eichenberg 22] では 60 億のパラメータを持つ GPT-J の重みを利用しているが，GPT-J は基本的に英語向けモデルであるため日本語に対する処理能力はあまり期待できない．そこで我々は 13 億パラメータを持つ日本語 GPT モデル [rinna Co., Ltd. 22] の重みを利用した．

3.2.4 Adapters

Adapters は Transformer の要素間に配置された一連のモジュールであり，パラメータの効率的な微調整の手法としてモデル重みの代わりに微調整することができる．ファインチューニングの際は Language Model(3.2.3) の GPT の重みは変化させず，Adapters モジュールのパラメータを学習させることによって大規模モデルのファインチューニングという高コストなタスクを少ない資源で実現する．

3.3 テキストプロンプト

3.1 にもある通り，提案システムではアイテム画像と共にテキストプロンプトによってカテゴリを指定することによって出力コメントの内容が変化する．例えばカテゴリがボトムスであるならば，入力画像のアイテムに対して「○○のようなボトムスが合う」といった内容のコメントが出力される．これによってモデルをカテゴリ数分だけ別に用意することなく，1 つのモデルのみで出力されるコメント内容を制御することを可能にした．

4. データセット

本研究で用いるデータセットは，実際にスタイリストが顧客に対して送った過去のコメントデータの一部である．このコメントデータを細かく分析し，推薦アイテムに合うアイテムについて言及している部分のみを抜き出し利用した．また，特に顧客情報に基づくパーソナライズされたコメントや季節，トレンドに関するコメントは除去した．最終的にコメントデータは約 94 万件になり，これを用いてファインチューニングを行った．

5. ファインチューニング

5.1 データについて

4 章のデータセットを用いて，MAGMA のファインチューニングを行う．チューニングに使う画像データはアイテム画像，キャプション (画像に対して付けられる説明文) はそのアイテムに対するコメントデータ，プロンプトはテンプレートベースで ”Q: {アイテム名} に合う {カテゴリ名} は何ですか？A:” の形をとる．これらの画像，キャプション，プロンプトの組みでチューニングを行った．

5.2 学習と損失

学習にあたって Language Model の重みは更新せず，Visual Encoder と Image Prefix，Adapters モジュールの重みを最適化する．言語モデルの構成要素は事前学習された日本語 GPT モデルの重みで初期化し，Visual Encoder は事前学習済み CLIP の重みを利用する．Image Prefix と Adapters は常に 0 から学習を行う．以下では学習可能なパラメータを添え字 θ で表す．学習タスクは画像キャプショニング (入力画像に対する尤もらしいキャプションを生成するタスク) であり，画像とキャプションのペア (x, y) が与えられるとそれぞれ

表 1: 提案システムのコメント生成例

図 3: 対象アイテム

カテゴリ	コメント例
トップス	ジャケットやカーディガンのインナーにも映えますね.
	ニットはウエストインしてネイビーやブラック，グレーなどのフレアスカートを合わせると，女性らしい着こなしに!
ボトムス	ボトムスには，ワイドパンツを合わせて大人カジュアルに，フレアスカートを合わせて女性らしい着こなしも素敵ですよ ^^
	チェック柄のスカートやレースのスカートに合わせて華やかにまとめるのもオススメです!
シューズ	靴はラウンドトゥパンプスで足元に抜け感を出すと◎
	足元には，グレージュ，こげ茶のショートブーツにも合わせやすいですよ.
アクセサリー	そのままサラリと着てもサマになりますが，デコルテに華奢なデザインのネックレスを合わせてクラスアップしても素敵ですね.
	アクセサリーはシルバーで統一して，足元はスエード素材のフラットシューズで，お楽しみいただくのも良いですね.

表 2: 定量評価結果

モデル	ROUGE-1	BLEU	FastText	BERT	MOVER
比較手法	0.433	12.57	0.880	0.759	0.583
提案システム	**0.483**	**17.91**	**0.885**	**0.821**	**0.619**

$$v_{1,\theta},\ldots,\, v_{n,\theta} = {V^p}_\theta \otimes {V^e}_\theta(x) \tag{1}$$

$$e_1,\ldots,\, e_m = E(t_1),\ldots,\, E(t_m) \tag{2}$$

のように埋め込み表現を獲得する．式 (1),(2) 中，V^p:Image Prefix ベクトル，V^e:Visual Encoder ベクトル，t_k: トークナイズキャプションを表す．なお画像列の長さ n は固定だが，キャプションの長さ m は可変である．このようにして埋め込まれたベクトル同士を連結し Language Model モジュールへと渡すことで，式 (3) によって損失を算出する．

$$L_\theta(x,\, y) = -\sum_{i=1}^{m} l_\theta(v_{1,\theta},\ldots,\, v_{n,\theta},\, e_1,\ldots,\, e_m) \tag{3}$$

式 (3) において l_θ は，Language Model モジュールの Transformer 出力をトークン空間に写像したベクトル H と，Adapters を組み込んだ Transformer モデルのベクトル $\tilde{T}$ を用いて $H \otimes \tilde{T}_\theta$ で計算される．

6. 実験

本章では，提案システム及び比較手法の精度を定量的，定性的の両観点から検証する．実験は，ファッションレンタルサービスのパーソナルスタイリングにおいてスタイリストが顧客に対してアイテムを推薦する状況で，アイテムに付け加える推薦コメントを作成する問題設定で行う．

6.1 比較手法

提案システムと比較する手法として，本研究では学習済み BERT[Jacob Devlin 19] と VisionTransformer[Alexey Dosovitskiy 21] を用いた対照学習による多ラベル分類モデルを採用する．このモデルは，アイテム画像とそれに対するコメントを入力としてラベルごとのスコアを出力する．そのスコアに基づき，最も近いコメントを 4 章のデータセットから引用することでコメント候補文とする．

6.2 候補コメントの生成例

ここでは，提案システムや比較手法がどのようなコメントを生成するのかについて具体的に取り上げる．例としてフィット感のあるハートネックニット (図 3) に対して生成されたコメントを，カテゴリごとに 2 文ずつ表 1 に示す．生成例を見るとカテゴリごとにコメント内容が正確に分かれており，かつ言葉遣いも安定している．このような生成結果を元に 6.3，6.4 にて評価を行う．

6.3 定量評価

本研究では提案モデルの定量的な評価をするにあたって，要約生成等のモデルの評価に広く利用される評価指標である ROUGE，機械翻訳等のモデルの評価に広く利用される BLEU，BERT によるエンベディングを利用した BERTScore[Tianyi Zhang 19]，文書間距離を測る手法である WMD を用いて BERTScore を改良した MoverScore の 5 つを用いた．さらに，ファッションデータ追加学習済み FastText によるエンベディングを利用した手法 (以降 FastText と表記) でも評価を行った．その結果について，スコアが上回った方を強調し表 2 に示す．

6.4 定性評価

本研究の実験を評価するにあたって，定量的な評価のみでは本質に至ることが難しい．これはファッションドメイン性や，正解が 1 つではないという生成タスク固有の問題が起因している．そこで本研究では，スタイリストによる人手評価を行った．アイテム画像とそれに対する提案モデルと比較手法の候補文を 3 文ずつ，さらに 4. のデータから抜粋した正解データ 3 文をカテゴリごとにランダムに並べ，それぞれに 0〜3 点をファッション評価と日本語評価の 2 つの観点でスコアリングする方式を採用する．この 2 つの観点によってファッションのドメインにおけるコメント内容の適切性と，生成文の構成や文法，言葉遣いに対する評価を明らかにすることができる．さらに，各カテゴリで最も良いと判断したものには最良文として最良ポイントを付与する．この方式に従い 3 名のスタイリスト

表 3: 定性評価結果

モデル	ファッション評価平均	日本語評価平均	累計最良ポイント平均
比較手法	2.197	**2.164**	**12.0**
提案システム	2.245	2.116	**12.0**
正解データ	**2.311**	2.147	9.67

によって 10 アイテム分評価し，その平均スコアを算出した結果を表 3 に示す.

6.5 考察

表 2 より，定量的には提案システムが比較手法をすべての指標で上回る結果になった．単語の意味を考慮せず，共通する単語に注目する ROUGE や BLEU と，単語の意味を考慮する BERT や MOVER，FastText の双方で上回っていることから，提案システムは比較手法よりも正解コメントの再現精度が高いと言える．

また表 3 の定性評価結果から，ファッション評価に関しては正解データが最も高いものの，提案システムは比較手法を上回っていることがわかる．この結果と定量評価の結果を踏まえると，提案システムは正解データと同じような単語・語彙を的中させつつ，少し異なった言い回しでコメントを生成できると考えられる．一方で日本語評価に関しては比較手法が最も高い結果となった．比較手法では出力コメントは生成されるのではなくデータセットから引用する仕様のため，コメント内容は人手で作成されたものである．その為，文を 0 から生成する提案システムよりも文の構成，言葉遣いなどの観点で上回り，このような結果になったと考えられる．

また表 3 の累計最良ポイント平均から，提案システムと比較手法どちらも正解データを上回る結果となった．ファッション評価では正解データが最も高いことを加味すると，比較手法や提案システムには評価が高い例と低い例の幅があると考えられる．

7. おわりに

本研究では，大規模モデルを Adapters ベースファインチューニングすることでファッションドメインへ適応させ，実運用するに十分な能力を持つシステムを構築した．特に定性評価の結果から，提案システムが人手と比較しても遜色ないものであることがわかる．さらに，ユーザ情報や時系列情報なども含めた学習をすることができればさらなるシステム価値の向上につながると考えられるので，今後の課題としたい．

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)

[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: equal technical contribution, equal advising: AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE(2021)

[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)

[Alec Radford 18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever: OpenAI: Improving Language Understanding by Generative Pre-Training(2018)

[Jacob Devlin 19] Jacob Devlin, Ming-Wei Chang , Kenton Lee , Kristina Toutanova: Google AI Language: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(2019)

[rinna.Co.,Ltd. 22] rinna Co.,Ltd. : https://huggingface.co/rinna/japanese-gpt-1b(2022)

[Tianyi Zhang 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)

[Ryotaro Shimizu 23] Ryotaro Shimizu , Yuki Saito , Megumi Matsutani , Masayuki Goto: Graduate School of Creative Science and Engineering, Waseda University,ZOZO Research,School of Creative Science and Engineering, Waseda University: Fashion intelligence system: An outfit interpretation utilizing images and rich abstract tags(2023)