参照画像と修正指示文を用いたMultimodal Modulationによるファッション画像検索

植田有咲† 杉浦孔明†
†慶應義塾大学
E-mail: arinko31@keio.jp

1 はじめに

E コマースの市場規模は年々拡大している. 特に, 日時や場所を問わずオンラインで物の売買ができることは便利であり, 人々の暮らしを豊かにしている. E コマースにおいて検索性能は重要であり, ユーザーが本当に欲しい商品を見つけるためにも高度な検索品質は必要不可欠である. しかしながら, 現状では商品検索においてユーザ意図を満たす商品を短いテキスト検索で見つけることは難しい. そのため, 本研究では商品検索において, 希望通りの商品が見つからない場合, 商品画像に対して修正指示文を用いることで検索品質を向上させることを目的として行っている.

本研究の対象タスクは Fashion 分野における画像再検索である. 本タスクは単純な画像検索タスクと異なり, 画像と修正指示文を元に新たな画像を再検索するタスクである. 元画像と言語情報には保持すべき情報と変更すべき情報が色やロゴ, 襟などが様々に存在し, 膨大な情報を見分け, 学習することは難しい. 実際に, 既存手法 [Lee 21] は形状などのコンテンツ情報, 色や柄などをスタイル情報を選択的に推論に用いることができるが, 実社会に適応するには性能が不十分である.

そこで本論文では Transformer を用いた画像再検索手法を提案する. 提案手法は Text Encoder, Source Image Encoder, Modulation Encoder, Candidate Image Encoder から構成される. 提案手法と既存手法の違いは CoSMo [Lee 21] における Contents Modulator と Style Modulaotor を代替するため, 参照画像と修正指示文を扱う Transformer [Vaswani 17] に基づく Modulation Encoder を導入したことである.

提案手法は Modulation Encoder を用いることで画像特徴量と修正文の各単語の関係性を学習することが可能なので, 性能向上が期待できる. 特に Modulation Encoder では Transformer の self-attention 構造を導入することで画像特徴量と修正文の各単語の関係性を学習することが可能である.

2 関連研究

修正指示文を用いた画像再検索手法として CoSMo, CIRPLANT [Liu 21], ARTEMIS [Delmas 22] がある. CoSMo はスタイル情報 (色, 柄等) とコンテンツ情報 (丈, 長さ等) を分離して学習を行うモデルであり, 提案手法のベースライン手法として用いている. ARTEMIS, CIRPLANT はそれぞれ Transformer を用いたモデルであり, 高い性能を記録している.

Fashion 分野の画像検索分野における標準データセットとして Fashion200K [Han 17], Fashion-MNIST [Xiao 17], Shoes [Berg 10] データセットがある. Fashion200K は様々なウェブサイトから収集された 200000 枚を超えるファッション画像と商品の説明情報, 服のバウンディングボックスから構成されている. Shoes データセットは 10000 枚の訓練集合と 4658 枚の検証集合から構成されており, 靴画像と商品の説明情報を含んだデータセットである. Fashion-MNIST データセットは 6 万枚の訓練集合と 1 万枚のテスト集合を含む計 7 万枚で構成されるデータセットである. MNIST 同様 10 種類のラベル付き画像データセットである. FashionIQ データセット [Wu 21] は約 8 万枚の服画像で構成されており, Dress, Toptee, Shirt の 3 種類のカテゴリに分割されている. データセット内にはペアとなる参照画像と対象画像が存在しており, それぞれのペアに対して 2 文ずつ修正指示文が与えられているデータセットである.

3 問題設定

本研究は Fashion 分野における修正指示文を用いた画像再検索タスクを対象としている. 本タスクでは参照画像とは異なる対象画像を修正指示文を用いて検索することを目的とする. 視覚言語情報を活用することで, よりユーザ意図を反映した検索結果を出力することが望ましい. 本研究の代表例を図 1 に示す. 図 1 のように薄い紫色の T シャツの参照画像に対して, “darker in color and more blue” といった修正指示文が与えられた場合に参照画像から対象画像のようなより暗い青色の T シャツを検索結果として出力することが望まし

参照画像　　対象画像

図 1　参照画像に対して “darker in color and more blue”という修正指示文が与えられた例

い. 提案手法の入力は参照画像, 候補画像群, 修正指示文である. 出力は画像と候補画像群をランキングした順序集合である.

本論文で使用する用語を以下に定義する.

- 参照画像: 修正を施す前の画像
- 対象画像: 参照画像とペアとなる 1 枚の正解画像
- 候補画像: カテゴリごとに存在する対象画像候補の画像
- 修正指示文: 参照画像から対象画像を検索するための指示文

本研究での評価尺度は Recall@k である.

4 提案手法

本手法は CoSMo 等の既存の Fashion 分野における画像検索手法と関連が深い. CoSMo を元となる手法とした理由は FashionIQ データセットにおいて良好な結果を得ているためである. 元となる手法との違いは CoSMo における Contents Modulator と Style Modulator を代替するため, 参照画像と修正指示文を扱う Transformer に基づく Modulation Encoder を導入したことである. 図 2 に提案手法のモデル図を定義する. 提案手法は Text Encoder, Source Image Encoder, Modulation Encoder, Candidate Image Encoder の 4 つのモジュールから構成される. 提案手法の入力は $x = \{\mathrm{x_{ref}}, \mathrm{x_{text}}, \mathrm{x_{cand}}\}$ である. $\mathrm{x_{ref}}$ は参照画像, $\mathrm{x_{text}}$ は修正指示文が複数ある場合は各文章を <SEP> トークンで結合させたものを表す. $\mathrm{x_{cand}}$ は候補画像を表す. $\mathrm{x_{ref}}$ を ResNet50 [He 16] の事前学習済みモデルの一部により構成された Source Image Encoder を用いて特徴量抽出を行った. $\mathrm{x_{text}}$ は BooksCorpus [Zhu 15] と English Wikipedia における事前学習済みの BERT [Devlin 19] で構成された Text Encoder を用いてトークンごとの埋め込みを行った. $\mathrm{x_{cand}}$ は Source Image Encoder とは異なり, ResNet50 の事前学習済みモデルの最終層までで構成された Candidate Image Encoder を用いて特徴量抽出を行った.

4.1 Feature Extractor

Source Image Encoder, Candidate Image Encoder ではそれぞれ $\mathrm{x_{ref}}$, $\mathrm{x_{cand}}$ を入力とし, ResNet50 を用いて特徴量抽出を行う. Text Encoder では BERT を用いて $\mathrm{x_{text}}$ から特徴量抽出を行う.

$$t = f_{text}(\mathrm{x_{text}}) \tag{1}$$

$$x^r = f_{img}(\mathrm{x_{ref}}) \tag{2}$$

$$v^t = f^*_{img}(\mathrm{x_{cand}}) \tag{3}$$

f_{text}, f_{img} は Text Encoder, Source Image Encoder を表す. f^*_{img} は Candidate Image Encoder を表す. ここで, Source Image Encoder からの出力は ResNet50 の conv5_x 層目からの出力であり, Candidate Image Encoder からの出力は ResNet50 の最終層からの出力である. $x^r \in R^{C\times(H\times W)}$, $t \in R^{L\times N}$, $v^t \in R^{512}$ であり, それぞれ符号化された参照画像特徴量, 修正指示文特徴量, 候補画像特徴量を表す. ここで C はチャネル数, H, W は画像の高さ, 幅を表す. L は文の単語数, N は Text Encoder の埋め込みベクトルの次元数である.

4.2 Modulation Encoder

Modulation Encoder の入力は x^r, t を結合したものである. 画像と言語特徴量を連結した後に Transformer に入力することにより, 画像特徴量と修正指示文の各単語の関係性を学習することが可能であると考えられる. x^r を畳み込み層に入力することで次元数を t の次元数に合わせた.x^r の畳み込み層からの出力を x_{img} とする. $x_{img} \in R^{1\times N}$ である. ここでチャネル方向に x_{img}, t の結合を行い, 画像言語特徴量 x_{vl} を得る.$x_{vl} \in R^{(L+1)\times N}$ x_{vl} に対し, Positional Encoding を行い, 位置情報の埋め込みを行う. Positional Encoding を行った x_{vl} を Multi-Head Attention 層に入力する. x_{vl} から Attention Head 数 A だけ Query Q, Key K, Value V を生成する.

$$Q^{(i)} = W_Q^{(i)} x_{vl}^{(i)} \tag{4}$$

$$K^{(i)} = W_K^{(i)} x_{vl}^{(i)} \tag{5}$$

$$V^{(i)} = W_V^{(i)} x_{vl}^{(i)} \tag{6}$$

ここで $i = \{1, ..., A\}$ である. 以下に示す Multi-Head Attention の式に基づき, Attention スコア S_{attn} を算出する. H は隠れ層のサイズを表す.

$$S_{attn} = f_{attn}^{(1)}, ..., f_{attn}^{(A)}, \tag{7}$$

$$f_{attn}^{(i)} = softmax(\frac{Q^{(i)} K^{(i)\top}}{\sqrt{d_k}}) V^{(i)} \tag{8}$$

$$d_k = \frac{H}{A} \tag{9}$$

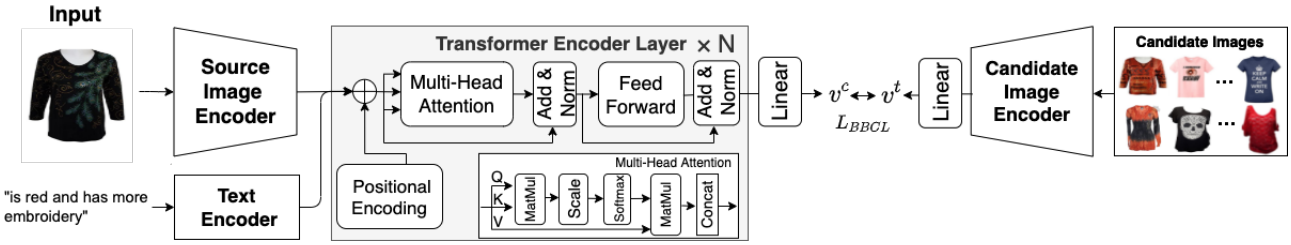

図 2　提案手法のモデル図

得られた S_{attn} はドロップアウト層, 正規化層を適用した後, 全結合層と活性化関数による処理が加えられる. 最後に再びドロップアウト層, 正規化層を適用する. この一連の処理を Transformer 層の 1 ブロックとして定義する. 最後の Transformer 層からの出力を h_{out} とする. h_{out} を全結合層に入力することで出力 $v^c \in R^{512}$ を得る.

4.3　損失関数

損失関数として batch-based classification loss (BBCL) を用いた. 損失関数の式を以下に示す.

$$L_{BBCL} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp \kappa(v^{c,i}, v^{t,i})}{\sum_{j=1}^{N} \exp \kappa(v^{c,i}, v^{t,j})} \quad (10)$$

ここに, v^c は提案手法の最終出力を表し, v^t は Candidate Image Encoder からの出力を表す. BBCL を用いて v^c と v^t の特徴量間の距離が最小となるように学習を行う. κ はコサイン類似度などの任意の類似度関数を表す.

5　実験設定

評価実験のデータセットには FashionIQ データセットを用いた. FashionIQ データセットは Fashion 分野における画像再検索タスクにおいて標準的に用いられている. FashionIQ データセットは 77684 枚の Fashion 商品画像から構成されており, Dress, Toptee, Shirt の三つのカテゴリに分類されている. 各サンプルにつき 1 枚の参照画像 x_{ref}, 対象画像 x_{trg}, 2 文の修正指示文 x_{text} で構成されている. 今回は Toptee のみを使用した. Toptee の総画像枚数は 26329 枚, 訓練集合は 5782 サンプル, テスト集合は 1847 サンプルである. すなわち, 候補画像 x_{cand} の画像枚数は 26329 枚, 参照画像 x_{ref} は訓練集合内に 5782 枚, テスト集合内に 1847 枚存在する. 修正指示文は Amazon Mechanical Turk によるものである. 訓練集合をモデルのパラメータの学習に, テスト集合をモデルの性能評価に使用した.

表 1 に提案手法の設定を示す. 表 1 の Transformer における#L は層数, #H は隠れ層の次元数, #A は Attention Head の数を表す. 提案手法の学習可能なパラメータ数は約 1 億 5000 万である. 学習は GeForce RTX 2080 Ti（メモリ 11GB）x1, 64GB-RAM, Intel Corei9 9900K を搭載した計算機上で行った. 学習に要した時間は約 2 時間半である.

表 1　提案手法のハイパーパラメータ設定

Backbone	ResNet50
Transformer	#L:8, #H:768, #A:8
Optimizer	RAdam ($\beta_1 = 0.9, \beta_2 = 0.999$)
Learning Rate	2.0×10^{-4}
Weight decay	5.0×10^{-5}
Batch size	32
Epochs	50

6　実験結果

6.1　定量的結果

本手法とベースライン手法の定量的比較結果を表 2 に示す. ベースライン手法として CoSMo [Lee 21] を用いた. CoSMo をベースライン手法として選んだ理由は FashionIQ データセットを含めた Fashion 系データセットにおいて良好な結果が報告されているためである. 評価尺度として標準的に用いられている Recall@k を用いた. Recall@k は以下で定義される.

$$\text{Recall@k} = \frac{\{\text{x}_{\text{trg}}\} \cap \text{P}_{\text{k}}}{\{\text{x}_{\text{trg}}\}} \quad (11)$$

ここで $\{x_{\text{trg}}\}$ は正解となる対象画像群, k は考慮する上位ランキング数, P_k は予測した上位 k 個の画像群を表す. ベースライン手法の Recall@10 は 21.07%, 提案手法は 22.32%であり, ベースライン手法に比べて 1.30%向上した. Recall@50, Recall@100 についても 44.70%から 46.40%, 57.75%から 58.71%へ向上した.

6.2　Ablation study

本研究の ablation 条件を以下に示す.

- Ablation-1 w/o BERT : ベースライン手法で用いられていた Text Encoder を用いた場合の性能への影響を調査するため, Text Encoder に BERT の代わりに LSTM を使用したモデル
- Ablation-2 w/ Transformer #L: 6 : Modulation Encoder の層数を減少させた場合の性能への影響

図 3　定性的結果：提案手法の成功例

図 4　定性的結果：提案手法の失敗例

表 2　定量的結果 Recall@k

Method	$k = 10$	$k = 50$	$k = 100$
CoSMo (reproduced)	21.07	44.70	57.75
Ablation-1 (ours)	20.64	44.27	55.86
Ablation-2 (ours)	21.79	45.23	58.12
Full (ours)	**22.31**	**46.40**	**59.39**

を調査するため Modulation Encoder の層数を 6 層に減らしたモデル

Text Encoder に LSTM を用いたモデルは提案手法に比べて約 2 ポイントほど性能が悪い結果が得られた．この結果より，Text Encoder として BERT を用いた単語ごとの埋め込みを行うことが性能向上に寄与していると考えられる．

6.3　定性的結果

提案手法の成功例を図 3 に示す．図 3 では一番左側の画像が参照画像，文章が修正指示文，右の緑枠または赤枠で囲まれた画像が予測結果の上位 5 枚の画像を表す．緑枠で囲まれた画像は正解画像を表し，予測結果は左にいくほど順位が高い．図 3 の上の例では鮮やかなピンクの T シャツに対して，“is brighter colored with no button”，“has more blue and less buttons” といった修正指示文が与えられることにより，提案手法は参照画像からより明るい青色のボタンのない T シャツを予測できていることが分かる．下の例については暗めの薄紫色の半袖の服に対して “The tank top is loose fitting and purple in color.”，“is sleeveless and is a brighter shade of purple” といった修正指示文が与えられることで参照画像から袖がない明るい紫色のタンクトップを予測できていることが分かる．どちらの例についても上位 1 位に正解画像を予測できている．

提案手法の失敗例はテストセットの中に合計 1400 サンプル存在した．提案手法が失敗した例を図 4 に示す．図 4 の一番左の画像は参照画像，水色の枠で囲まれた画像は正解画像，文章が修正指示文，右の赤枠の画像が提案手法の予測結果の上位 5 枚である．成功例と同様予測結果は左にいくほど順位が上位の画像である．“sleeves”，“is black with skull designs” という修正指示文が与えられているが，提案手法は袖無しのタンクトップを予測結果として出力している．失敗した原因としては修正指示文の一つが “sleeves” のみであり，正解例の修正指示文に比べて情報量が少ないため，袖情報が重視されなかったためと考えられる．

7　結論

本論文では Fashion 分野における画像再検索タスクに焦点を当てた．本論文の貢献は以下の通りである．

- CoSMo における Contents modulator と Style modulator を代替するため、参照画像と修正指示文を扱うに基づく Modulation encoder を導入した．
- FashionIQ データセットの Toptee カテゴリにおいて既存手法に比べて性能を向上させた．

将来研究としては性能向上のために骨格検出の導入や Optical Character Recognition (OCR) などの服上の文字についても考慮した予測を行うことが可能であると考えられる．

謝辞

本研究の一部は，JSPS 科研費 20H04269, JSTCREST, NEDO の助成を受けて実施されたものである．

参考文献

[Berg 10] Berg, T. L., Berg, A. C., and Shih, J.: Automatic attribute discovery and characterization from noisy web data, in *ECCV*, pp. 663–676Springer (2010)

[Delmas 22] Delmas, G., Rezende, de R. S., Csurka, G., and Larlus, D.: ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity, *arXiv preprint arXiv:2203.08101* (2022)

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, pp. 4171–4186 (2019)

[Han 17] Han, X., Wu, Z., Huang, P. X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., and Davis, L. S.: Automatic spatially-aware fashion concept discovery, in *ICCV*, pp. 1463–1471 (2017)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in *CVPR*, pp. 770–778 (2016)

[Lee 21] Lee, S., Kim, D., and Han, B.: CoSMo: Content-style modulation for image retrieval with text feedback, in *CVPR*, pp. 802–812 (2021)

[Liu 21] Liu, Z., Rodriguez-Opazo, C., Teney, D., and Gould, S.: Image Retrieval on Real-Life Images With Pre-Trained Vision-and-Language Models, in *ICCV*, pp. 2125–2134 (2021)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, *NeurIPS*, Vol. 30, (2017)

[Wu 21] Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., and Feris, R.: Fashion IQ: A new dataset towards retrieving images by natural language feedback, in *CVPR*, pp. 11307–11317 (2021)

[Xiao 17] Xiao, H., Rasul, K., and Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, *arXiv preprint arXiv:1708.07747* (2017)

[Zhu 15] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in *ICCV*, pp. 19–27 (2015)