D-001

画像認識によるファッションアイテム推薦システムの構築

Building a fashion item recommendation system using image recognition

竹内 惟織† 土屋 誠司‡ 渡部 広一‡
Iori Takeuchi Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，好きなキャラクターやアイドルなどへの熱狂的な応援活動として「推し活」が注目されている．推し活はファッションにおいても展開されており，推しのアイドルや芸能人が着用している衣類や帽子，靴などのファッションアイテムを真似し，推しと同じ気分を味わったり，推しの着こなし方やコーディネートが参考になることもある．

しかし，推しが芸能人などであると，着用アイテムが高価なブランド品やレアなアイテムであることが多々あるため，これらを入手するには困難な場合がある．そこで，推しの着用アイテムに類似したファッションアイテムを検索することができれば，高価なものが購入できなくても推しと同じ気分を味わえるのではないかと考える．

本研究では，推しが身に付けているファッションアイテムに類似したアイテムを推薦するシステムの構築を目的とする．手法として，Deepfashion2[1]というファッションに関するデータセット用いて You Only Look Once (YOLO)[2]でファッションアイテムの検出ができるように学習モデルの作成を行う．類似度学習には Triplet Loss[3]を使用して，データベース内の画像と入力画像の特徴抽出には VGG16[4]を用いる．それを基にコサイン類似度を計算し，類似度の高い順にアイテムを推薦する．

2. 関連技術

2.1. YOLO (You Only Look Once)

本研究では You Only Look Once[2](以降 YOLO)と呼ばれる物体検出アルゴリズムを採用する．YOLO は，画像中の物体を高確率で検出するための深層学習ベースのアルゴリズムであり，一度の処理で物体検出を行う．

YOLO の特徴は，画像全体を一つのグリッドに分割して，各グリッドで物体の存在を同時に予測することである．これにより，他の物体検出アルゴリズムよりも高速に高い精度で検出することができる．各グリッドセルでは，複数のアンカーボックスを使用して異なるスケールやアスペクト比の物体を同時に検出すことが可能となっている．

2.2. Triplet Loss

類似度学習として，Triplet Loss[3]を使用する．これは，損失関数の一種で，画像検索などのタスクにおいて，データの埋め込みを学習するために利用されるアルゴリズムである．正解画像の Anchor，それと同じカテゴリの Positive 画像，異なるカテゴリの Negative 画像の 3 つを 1 組で入力とし，Anchor 画像と Positive 画像の距離が Anchor 画像と Negative 画像の距離よりも小さくなるように学習を行う．図 1 に，Triplet Loss のイメージ図を示す．

2.3. VGG16 (Visual Geometry Group 16-layer)

特徴抽出のアルゴリズムとして，VGG16 (Visual Geometry Group 16-layer)を使用する．VGG16 は， 16 層からなる深い畳み込みニューラルネットワークで構成されており，事前学習済みの重みを使用して特徴抽出を行うことが可能である．

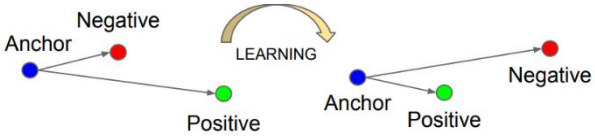

図 1 Triplet Loss のイメージ図

3. 提案手法

3.1. データセット

本研究では，YOLO の学習に「DeepFashion2」，商品データとして，楽天 API を利用した楽天市場のデータを使用する．

3.1.1. DeepFashion2

DeepFashion2[1]は，ファッションに関する大規模なデータセットで，構造は，13 の異なるファッションカテゴリーが含まれており，ファッションを着用した人々の画像，バウンディングボックスの位置や衣服の領域を示したアノテーションデータなどから構成されている．このアノテーションデータを基に学習させ，YOLO でファッションアイテムの検出ができることを目指す．

3.1.2. 楽天データ

楽天 API を使用し，ファッションに関するキーワード（例：メンズファッションなど）を指定し，商品情報，その商品に付随するタグを獲得し保存する．これを繰り返し行うことで商品データセットを作成する．このデータを商品データ，および類似度学習，評価に用いる．

3.2. 学習

ファッションアイテムを YOLO で検出するためには，学習をさせてファッション用のモデルを構築する必要がある．また，YOLO の学習モデルに合わせるため，DeepFashion2 から抽出したバウンディングボックスの座標を中心座標，ボックスの高さ，幅に変更する必要がある．そのため DeepFashion2 のアノテーションデータから，カテゴリ ID とファッションアイテムの領域を示すバウンディングボックスの座標を抽出し，それらを基に学習を行う．

類似度学習では，VGG16 の事前学習済みモデルを Triplet Loss を用いて転移学習を行う．Anchor 画像は，ランダムに選択し，Positive 画像，Negative 画像はそれぞれ，Anchor 画像とのタグの類似度によって決められる．類似度を求める際に使用するのは，Jaccard 類似度で，2 つのセットの交差部分のサイズを和集合のサイズで割った値を基に，最も類似度が高いものを Positive 画像に，類似度が低いものを Negative 画像として扱う．また，目的関数のパラメータとして，ユークリッド距離を使用して距離計算を行う．

† 同志社大学大学院理工学研究科
‡ 同志社大学理工学部インテリジェント情報工学科

3.3. アイテム検出と推薦

学習したモデルを使用して，YOLO で入力画像中にあるファッションアイテムを検出し，その領域に基づいて画像をアイテムごとに切り取る．切り取られたアイテム画像と商品データ画像それぞれに対して，Triplet Loss で転移学習したモデルを使用して，特徴抽出をして，類似アイテムを推薦する．

4. 評価手法

YOLO の学習に対する評価は，定量的な手法としては再現率，適合率を用いる．再現率とは，あるグループが存在するときに，そのグループ内のデータをモデルがどれだけ見逃さずに検出できたかを示す指標で，適合率は，モデルがあるグループと予測したデータの中で，実際に正解であるものがどれだけ含まれていたかを示す指標のことである．

類似推薦に対する評価は，入力画像と推薦された画像のタグの一致数を求め，その一致数がデータセット全体をタグの一致数順に並べた際に何番目に位置するのかというので精度を評価する． つまり，推薦画像が全体でのタグ一致数による順位でも高ければ，精度が良いと考える．

5. 結果および評価

YOLO の学習結果は， 適合率，再現率が約 76%，モデル全体的な性能を評価した平均適合率が 80%という結果を得た．この学習したモデルを使用した，ある入力画像に対する物体検出すなわち YOLO の物体検出の結果が図 2 である．左側は長袖 T シャツとズボン，右側は半袖とズボンをそれぞれ領域とカテゴリの検出を正しく行えている．

図 2　YOLO の物体検出結果

図 3 は，事前学習済みモデルを使用した VGG16 による推薦結果で，推薦画像が高い割合でランキング上位に来ていることがわかる．図 4 は，タグを基に VGG16 を転移学習させた推薦結果で， 転移学習なしの VGG16 よりも下位の割合が高い結果となった．このことから，タグを基に転移学習を行ったモデルの精度が低いことがわかる．

6. 考察

まず，YOLO の検出について，現在データセットの中には，背景が多く映り込んでいる画像やカテゴリに偏りがあるものがある．背景が多く映り込んでいると，特徴抽出の際にノイズとなりえる可能性がある．カテゴリの偏りは，半袖やズボンなどよく着られているもののデータは多く，長袖ドレスなどは少なくなっている．よってデータセットの調整を行うことで，ノイズが軽減され，よりバランス良く学習できるためより精度を上げることが期待される．

類似推薦においては，タグが一致していれば，その画像同士は似ているであろうという考えの下，タグ一致による学習と評価を行ったが，タグが多く一致していても，画像自体が類似していないことがあった．これは，同じ店舗の商品だと，たとえ商品が異なっていても，似たようなタグが付けられていることが原因だとわかった．Triplet Loss における学習では，単にタグの類似度だけで Positive, Negative を決定するのではなく，画像自体の特徴量も考慮して決定していくことが必要と考える．評価においては，タグ間の距離（ベクトル化）からの評価などを組み合わせると，より詳細な評価がとれるのではないかと考える．

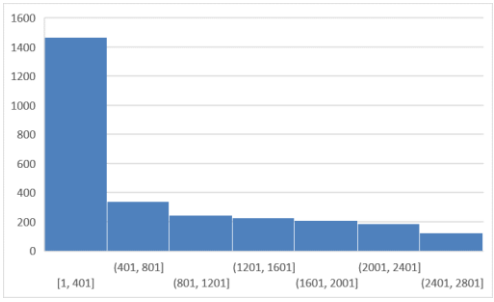

図 3　タグの一致数による全データの順位分布（VGG16）

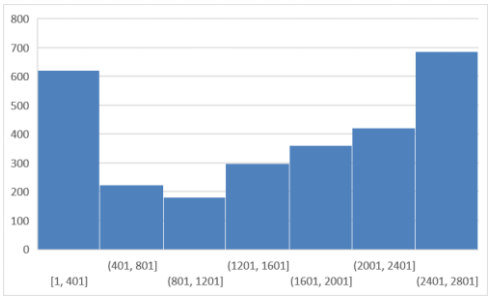

図 4　タグの一致数による全データの順位分布（転移モデル）

7. おわりに

本研究では，画像認識によるファッションアイテム推薦システムの構築を行った．システムは，構築することができたが，精度に関しては 課題が残る結果となった．特に類似度学習の部分では，タグの類似度による学習の精度である．この点については，全ての画像のタグに，理想的なもの（カテゴリ名，色，性別など）が含まれているのか， 含まれていなければ，タグの自動付与などの検討も必要になってくる．

今後の展望として，YOLO による検出でカバンや靴などのカテゴリ追加や現在は，性別関係なく類似度の高いものを推薦するシステムとなっているが，女性用のものを男性バージョンに変換して推薦したりなどファッションに関する推し活に特化したシステム開発を目指す．

参考文献

[1] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In CVPR, pp. 1096–1104, 2016.
[2] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. arXiv. “You Only Look Once: Unified, Real-Time Object Detection”. *arXiv preprint arXiv:1506.02640*, 2015. 6
[3] Florian Schroff, Dmitry Kalenichenko, James Philbin. arXiv. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. arXiv preprint arXiv:1503.03832, 2015. 6
[4] Karen Simoyan, Andrew Zisserman. arXiv. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. *arXiv preprint arXiv:1409.1556*, 2015.10