意味と形状の分離による マルチモーダルレシピ検索及び画像生成

杉山 優[1,a)]　岡本 開夢[1,b)]　柳井 啓司[1,c)]

概要

レシピのマルチモーダル検索タスクの研究の多くでは、テキストと画像を同じ意味空間にエンベッディングすることで検索を行う。本研究では画像のエンコードを意味と形状の二つに分離することで、検索に有用な情報のみを検索に使用する RDEGAN(Recipe Disentangled Embedding GAN) を提案した。マルチモーダル検索タスクにおいて初めて意味と形状の分離を実現し、Recipe1M [6] 検索タスクにおいてをこれまでの最高精度を達成し、さらに生成画像の品質も向上させることに成功した。

1. はじめに

昨今のウェブサービスの発展により，人の暮らしにおいてインターネットを通してレシピを検索したり，推薦されたりすることが広まっている．Cookpad や AllRecipes などのサービスを利用してユーザーはレシピテキストと画像を含む投稿を行うことができる．この多くのデータは各サービスのフォーマットに則って作成されているため，スクレイピングを行うことで大規模なデータセットを作成することができる．

Recipe1M [6] データセットでは 12 のレシピ投稿サイトから料理画像と調理手順や材料を含むテキストのデータを収集した．この大規模データセットの登場によって，テキストと画像のマルチモーダル学習を行う研究が盛んに行われるようになった．

Joint Embedding [7] や AdaMine [1] ではテキストと画像を同じ形のベクトルにエンコードし，その潜在空間内でのテキスト意味ベクトルと画像意味ベクトル間のユークリッド距離を最小化することによって共有空間を学習し，両意味ベクトル間の距離によって検索を行った．

R2GAN [9] や ACME [8] ではエンコードするだけでなく，共有空間からの画像やテキストの再構成を行うことによって共有空間での意味の表現がより高レベルな情報を含むようになり，検索精度が改善することを示した．

しかしこれらの手法では画像の検索と生成を全く同じ特徴量から行っており，異なるタスクでの中間表現の違いについて考慮されていない．レシピ画像にはテキストで表現される意味的な特徴である具材の色や形などの情報と，画像のみでしか表現されない盛り付けの形や皿の形状，カメラの画角などの情報が両方含まれている．もしこれらを分離し，テキストと画像で共有される情報を検索に使用し，画像のみでしか使用しない情報を画像生成に使用することができれば、検索と画像生成の両方で性能向上が期待できる．

本研究では，マルチモーダルレシピ検索タスクに対して，テキスト意味ベクトルと画像意味ベクトルの両者を共有空間へとエンベッディングを行う RDEGAN を提案する．この際，画像を検索に使用する意味ベクトルと画像再構成に使用する形状特徴の二つにエンコードすることによって，画像からの検索と画像からの再構成のそれぞれのタスクに適応した特徴量を学習し，検索精度と画像生成のどちらにおいても品質を向上させることが目的である．この提案により、マルチモーダル検索タスクにおいてはじめて画像の形状と意味の分離を行い、Recipe1M 検索ベンチマークでこれまでの最高精度を上回る精度を達成し、さらに画像生成の品質も向上させた。

2. 関連手法

2.1 マルチモーダルレシピ検索

初めて Recipe1M [6] を使用するマルチモーダル画像検索を提案したのは JE(Joint Embedding) を用いる im2recipe [7] で，画像とテキストをそれぞれエンコードし，それらをコサイン類似度の尺度で最適化することで共有空間を学習した．

現在の state-of-the-art な手法である ACME [8] では，マルチモーダルレシピ検索の改良手法としてエンベッディングから画像の再構成を行い，潜在空間ではトリプレット学習を行うものを提案した．画像の生成によって検索精度が向上する手法に加え，テキストと画像から作られた検索用

1 電気通信大学大学院情報理工学系研究科情報学専攻
a) sugiya-y@mm.inf.uec.ac.jp
b) okamoto-ka@mm.inf.uec.ac.jp
c) yanai@cs.uec.ac.jp

ベクトルが，それぞれどちらから生成されたものなのかを判別しづらくするような仕組みを追加した．さらに，画像の再構成だけではなく，テキストの再構成としてレシピに含まれていた材料と，料理のカテゴリに対してそれぞれのクラス分類問題を検索用エンベッディングから行うことで，より精度を向上させた．

2.2 画像の形状と意味の分離

画像の生成を行う際に，潜在空間の解釈性に欠ける問題点に対して，画像の意味的な特徴と形状的な特徴を分離するという手法の研究が行われている．DRIT [4] では意味と形状の分離のためにそれぞれのエンコーダを用意し，一つの画像を 2 つのエンコーダに通して意味特徴と形状特徴を抽出し，他の入力画像から抽出した特徴と入れ替えて画像の再構成を行い，再構成された画像を再びエンコーダに入力した際の意味特徴と形状特徴が元画像のエンコード結果と近づくように学習を行うことで，意味と形状の分離を行った．MUNIT [3] では，意味特徴の抽出の際，空間的な特徴量を潰すためにグローバルアベレージプーリングを行い，形状特徴と合わせて画像再構成を行った．

どの手法においても，形状と意味のエンコードには異なるエンコーダを用いて，再構成された画像をエンコードした結果が元のエンコードと同一になるよう学習する点で共通しており，本研究での形状と意味の分離においてもこれを踏襲したアーキテクチャを用いる．

3. 手法

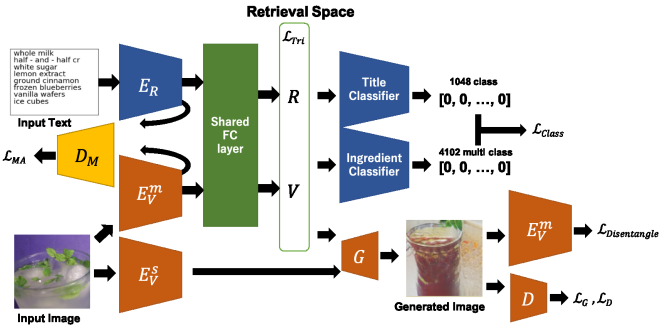

図 1 本手法 RDEGAN でのネットワークの概要図.

本手法の概要を図 1 に示す．この概要図は最終的なネットワークの概要であり、後述する第二段階の学習を行った場合の概要図である。本提案では ACME など既存のマルチモーダル検索モデルと同様に，テキストエンコーダ E_R と画像の意味エンコーダ E_V^m と重みを共有された全結合層 **FC** により共有空間へエンベッディングを行い，テキスト意味エンベッディング $R = \mathbf{FC}(E_T(\mathbf{t}))$ と画像意味エンベッディング $V = \mathbf{FC}(E_I^m(\mathbf{i}))$ の両者から画像の生成とクラス分類を行う．テキストエンコーダの出力と画像意味エンコーダの出力は，対応するテキストと画像の場合同じになるべきであるため，これらを近づけるためのディスクリミネータ D_M を使用して敵対的に学習する．画像生成の際にはレシピの意味を示す R と V のどちらかと画像形状エンコーダ E_V^s の出力を組み合わせて画像を生成することで，意味と形状の両方を利用した生成を行い，ディスクリミネータ D と対立的に学習することで画像生成を学習する．

テキストと画像のエンコーダにより共有空間に写像を行い，そのエンベッディングから画像の生成とクラス分類を行う．この際に画像をエンベッディングに使用するエンコーダと，画像生成に使用するエンコーダの二種でそれぞれエンコードする．

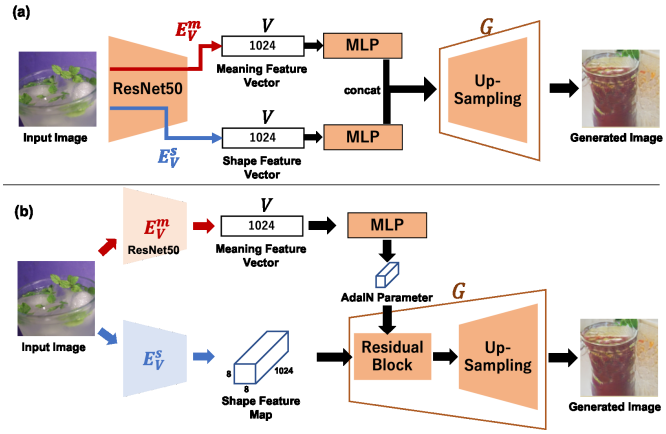

図 2 (a) は Encode Vector での，(b) は Encode Map での画像エンコーダとデコーダ.

3.1 二種類の画像エンコード

本手法での画像エンコーダの学習は二段階に分けて学習し，それぞれの概要を図 2 に示す．第一段階 (Encode Vector) では画像の検索に形状的な画像の生成に使用するような特徴を使用しないことを学習し，第二段階 (Encode Map) で分離された特徴を使用した画像生成を学習する．

画像のエンコーダの形状に合わせて画像生成器のアーキテクチャも変化しており，それぞれ学習する．画像の意味特徴をベクトルで表現する手法 (Encode Vector) では，意味特徴と形状特徴を同じエンコーダから同じ形状でエンコードする．この第一段階の学習の目的は，意味エンコーダが低レイヤの特徴である形状特徴を使用せずにレシピの意味だけを表現するように学習することで，画像の意味が十分に表現され，検索精度を向上させる．

画像の意味特徴を特徴マップで表現する手法 (Encode Map) では，Encode Vector の学習で作成された形状を分離した意味特徴エンベッディングから画像を生成することを学習する．特徴マップは空間的で，形状的な特徴を示すため，これと意味をあらわすエンベッディングを組み合わ

せて学習することで意味と形状を分離した画像生成を学習する．この形状エンコーダと画像生成器を学習することで，意味と形状を分離して表現する画像生成を行えるようにネットワークを学習することが第二段階の学習の目的である．

第一段階では E_V^m, E_V^s の初期値に ImageNet で事前学習した重みを用る。第一段階から第二段階に移行する際には E_V^m と重み共有の全結合層 **FC** を第一段階で学習たものを用い、ほかのネットワークはランダムな初期値から学習を行う。

3.2 画像とテキストの共有空間へのエンベッディング

共有空間へのエンベッディングは ACME と同様に行う．まず画像とテキストの両者を 1024 次元の意味ベクトルにまでエンコードする．画像のエンコードには ResNet50 の最終層を除いた ImageNet で訓練済みのモデルに全結合層を追加したモデルを使用する．これによってエンコードされた画像の特徴を意味特徴として利用する．テキストのエンコードは Recipe1M や ACME と同様に材料リストを双方向性 LSTM，調理手順を階層 LSTM で意味ベクトルまでエンコードしたものを使用する．

エンコードされた二つの意味ベクトルを，重みを共有した全結合層 **FC** を通すことで共有空間へのエンベッディングを行う．共有空間への写像を行う際に，二つのドメインのベクトルを重み共有した層に入力することにより，共有空間内でベクトルの距離学習することが直接意味同士の距離を最小化することにつながる．

3.3 画像の意味と形状を分離した生成

画像の生成の学習を行う際、第一段階の Encode Vector、第二段階 Encode Map のどちらの手法においても他画像生成の手法と比較して，実験的に精度が高く、学習に時間がかからないことがわかったので，LSGAN [5] を利用する．

画像を意味的にエンコードするだけではなく，形状的な特徴もエンコードするようにエンコーダをそれぞれ用意し，意味的な特徴を重み共有レイヤに入力する．画像の意味特徴はレシピ分類問題を解くことでレシピの意味とテキストとの距離を学習を行う．形状特徴は画像生成の際のみに利用し，レシピの検索には使用しないが画像の生成には必要な特徴として利用する．

異なる意味特徴や形状特徴を持つように生成した画像と実画像をそれぞれエンコーダに入れた出力が近づくような損失を，MUNIT で使用されていた損失を採用して，L1 距離によって最適化を行う．

3.4 学習プロセス

全体のネットワークの損失を GAN の学習過程で最小化することで全体の学習を行う．ただし，学習はまず Encode Vector で学習を行い，その学習で得られる訓練済みの画像とテキストの意味エンコーダ E_R と E_V^m のパラメータを固定し，次に Encode Map で学習を行う．この学習過程で得たモデルを提案手法である RDEGAN の学習済みモデルとする．この学習過程をとる理由は，最初から形状特徴を特徴マップとした手法を学習しようとすると，生成画像を特徴マップのみからある程度の精度で再構成できてしまうため，意味特徴ベクトルがノイズだったとしても損失関数が下がり，レシピ検索のための表現をよく学習することができなくなってしまうことが実験的にわかっているためである．

4. 実験

本実験に使用するモデルは，Recipe1M [6] データセットの学習データセットで学習し，テストデータセットで評価した．

4.1 検索の評価

検索を行った例を図 3 に示す．バタークッキーに関するテキストを入力すると，バターや砂糖などに関係の深いバタークッキーの画像やケーキなどの画像が検索上位に挙げられ，サラダの画像を入力すると材料リストに葉やトマトなどが含まれているテキストが上位に挙げられた．

次に，定量評価を行った．既存手法と提案手法でレシピ検索の精度がどれだけ変化したかを MedR とリコール率で評価した．MedR は全データの検索順位を良い順に並べ，その中央値を表す指標で，リコール率は全データの検索の内，ある順位以内に正解データが存在した割合を示す指標である．この結果を表 1 に示す．既存の手法よりも提案手法が高い精度で検索が行えており，特に 10,000 サンプルでの実験で精度を大きく向上させることができた．

表 1 本手法と既存手法の検索精度比較.

#		画像からレシピ			レシピから画像		
	手法	MedR↓	R@1↑	R@5↑	MedR↓	R@1↑	R@5↑
1k	JE	5.2	25.6	51.0	5.1	25.0	52.0
	R2GAN	2.0	39.1	71.0	2.0	40.6	72.6
	ACME	**1.0**	51.8	80.2	**1.0**	52.8	80.2
	RDEGAN	**1.0**	**59.4**	**81.0**	**1.0**	**61.2**	**81.0**
10k	JE	41.9	-	-	-	-	-
	R2GAN	13.9	13.5	33.5	12.6	14.2	35.0
	ACME	6.7	22.9	46.8	6.0	24.4	47.9
	RDEGAN	**3.5**	**36.0**	**56.1**	**3.0**	**38.2**	**57.7**

4.2 画像生成の評価

生成画像の例を図 4 に示す．提案手法で既存手法の ACME [8] よりも入力画像に近い画像を再構成することができた．

生成画像がどれだけもっともらしく作られたかの評価を，FID [2] を用いて ACME と比較した．FID というのは，元

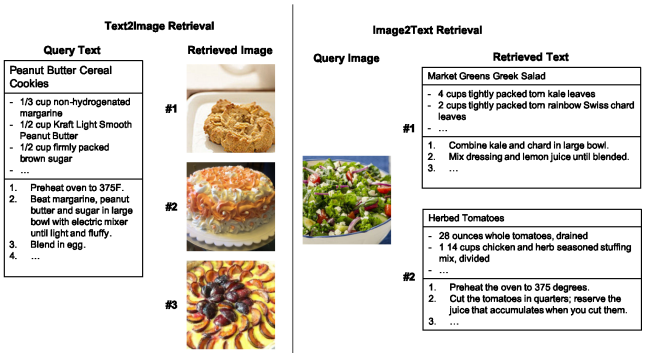

図 3　テキストからの画像検索 (左) と画像からのテキスト検索 (右) の一例.

表 2　本手法と既存手法の生成画像の FID 比較.

手法	FID@画像から再構成 ↓	FID@テキストから再構成 ↓
ACME	183.8	182.9
EV(ours)	162.8	168.2
EV+EM(ours)	**158.9**	**158.6**

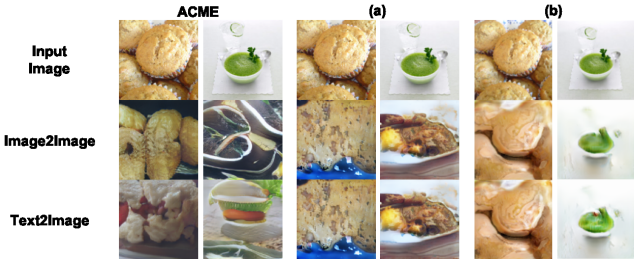

図 4　既存手法と提案手法の，Encode Vector の生成画像 (a) と Encode Vector + Encode Map(RDEGAN) の生成画像の比較 (b)．Image2Image は画像から生成した embedding からの，Text2Image はテキストから生成した embedding からの，それぞれ画像生成結果を表す.

の分布と生成した画像の分布がどれだけ近いかを示す尺度で，つまりどれだけもっともらしい画像を幅広く生成したかを示す．本手法と既存手法で，Encode Vector(EV) での生成画像と Encode Map(EM) での生成画像の分布について，Recipe1M データセットの画像との分布を計測したものを表 2 に示す．全ての場合において，提案手法が最も元の分布に近い画像を生成したことが検証できた.

4.3　画像特徴の表現空間の評価

本手法での画像生成は形状特徴を変化させることで画像生成生成結果を含まれる材料などの意味を保ちつつ変化させることができる.

意味特徴と形状特徴をそれぞれ連続的に変化させた際の生成画像の変化について実験を図 5 に示す．意味特徴を変化させた際には色やテクスチャなどの画像的な変化が，形状特徴を変化させた際には背景や皿や盛り付けなどの情報が変化することがわかる.

5.　おわりに

本手法では，既存のレシピ検索ではレシピの意味とは異

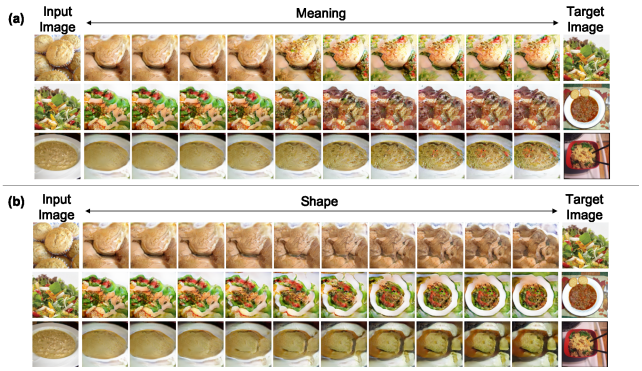

図 5　(a) は意味特徴ベクトルを連続的に変化させた際の生成画像の変化，(b) は形状特徴マップを連続的に変化させた際の生成画像の変化.

なる情報を意味と同列に扱ってしまっていた問題を解決するために，マルチモーダルレシピ検索タスクにおいてはじめて画像の意味と形状を分離することで検索精度と画像生成品質を改善する手法である RDEGAN を提案した．レシピ画像の情報を検索に使用する部分と使用しない部分で分離することが検索精度の向上に役立つことを示し，Recipe1M データセットでの検索精度において新しい state-of-the-art を達成した．将来的にはさらなる画像生成品質の向上や学習過程の簡易化などを検討している.

謝辞　本研究は JSPS 科研費 15H05915, 17H01745, 19H04929, 17H06100 の助成を受けたものです.

参考文献

[1] Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N. and Cord, M.: Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings, *ACM SIGIR*, pp. 35–44 (2018).
[2] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, *NeurIPS* (2017).
[3] Huang, X., Liu, M., Belongie, S. and Kautz, J.: Multimodal Unsupervised Image-to-Image Translation, *ECCV*, pp. 172–189 (2018).
[4] Lee, H., Tseng, H., Huang, J., Singh, M. and Yang, M.: Diverse Image-to-Image Translation via Disentangled Representations, *ECCV*, pp. 35–51 (2018).
[5] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z. and Paul, S. S.: Least squares generative adversarial networks, *ICCV*, pp. 2794–2802 (2017).
[6] Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I. and Torralba, A.: Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images, *IEEE T-PAMI* (2019).
[7] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I. and Torralba, A.: Learning Cross-modal Embeddings for Cooking Recipes and Food Images, *CVPR* (2017).
[8] Wang, H., Sahoo, D., Liu, C., Lim, E. and CH, Hoi, S.: Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images, *CVPR* (2019).
[9] Zhu, B.and Ngo, C., Chen, J. and Hao, Y.: R2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network, *CVPR*, pp. 11477–11486 (2019).