論文

重要語に着目した写真列からのレシピの自動生成

西村 太一[†]・橋本 敦史[††]・森 信介[†††]

本研究では，写真付きレシピの作成を容易にするために，写真列を入力としてレシピを生成するという課題と，それを解決する手法を提案する．レシピを正しく生成するためには，モデルは写真を説明する上で欠かせない物体や動作といった重要語と，それを含む表現を生成する必要がある．写真列を入力として文章を出力する類似課題である Visual storytelling の手法では，重要語の存在は考慮されていなかった．これに対して，本論文では，検索課題として取り組まれてきた手法を文生成モデルに組み込むことで，モデルは入力写真に適した重要語を過不足なく含む表現の情報を活用しながらレシピを生成する手法を提案する．日本語のレシピを対象に実験を行なった結果，本手法を適用することで生成文の自動評価尺度や，写真に適した重要語が生成文中に含まれているかといった評価においてベースラインと比較して性能が向上したことを実験的に確認した．

キーワード：レシピ，写真列，共有潜在空間，文生成

Recipe Generation from a Photo Sequence by Focusing on Verbalizing Important Terms

Taichi Nishimura[†], Atsushi Hashimoto[††] and Shinsuke Mori[†††]

This paper proposes a new problem for generating recipes from photo sequences and suggests a new method to more successfully achieve this, which aims to help users obtain multimedia recipes only by taking photographs. For this purpose, the output texts should include expressions with important terms that make sense as instructions. However, traditional methods proposed in “Visual storytelling” do not consider these expressions. To select expressions with important terms to describe a photo, the proposed method incorporates a retrieval method as well as a generation model. The proposed method was implemented and tested using Japanese cooking recipes. From various experimental results, it was confirmed that the new method outperforms standard baselines.

Key Words: *Recipe, Photo sequence, Cross modal embedding, Sentence generation*

[†] 京都大学大学院 情報学研究科, Graduate School of Informatics, Kyoto University
[††] オムロン サイニックエックス株式会社, OMRON SINIC X Corporation
[†††] 京都大学学術情報メディアセンター, Academic Center for Computing and Media Studies, Kyoto University

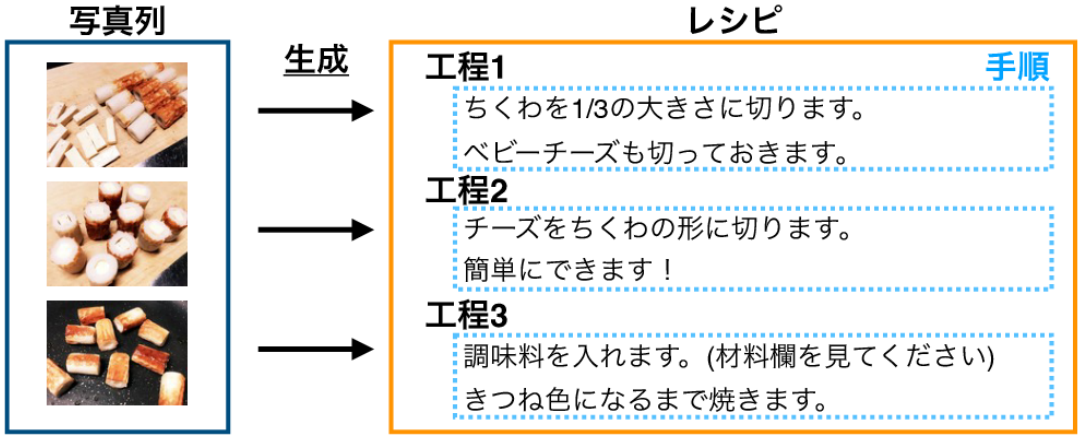

図 1　写真列からのレシピの自動生成．入力が写真列であり (左)，出力が複文からなる手順である (右)．手順は写真列の各写真ごとに生成する．

1　はじめに

言語による指示に加えて，その指示内容を示す動作途中の写真があれば，その写真を参考にして調理を行いやすくなる．したがって，各手順に写真が付与された「写真付きレシピ」により作業内容を示すことは有益である．しかし，写真付きレシピを作成するためには，写真を撮影しながら手順を実施し，実施後に各写真に対応する手順を記述する必要があり，作者にとって負担である．本研究の目的は，写真列を入力としてレシピを自動生成することで，写真付きレシピの作成を容易にすることである．この目的を達成するために，本論文では，写真列を入力として与え，システムは各写真ごとに手順を生成する問題として定式化した課題と，この課題を解決する手法を提案する．

図 1 に本論文で対象とする課題の概要を示す．入力の写真列の各写真に対し，複数の文からなる手順が対応している．これらの手順全体を本論文ではレシピと呼ぶ．本論文で取り上げる写真列の各写真は手順実施の上で重要な場面で写真を撮影したものであり，手順途中の情報が十分に含まれている．また，各写真に対して 1 つの手順が対応するため，生成すべき手順数が既知である．システムはこの写真列を受け取り，写真列の各写真に対応する手順を生成し，それらをまとめて写真付きレシピとして出力する．この課題設定は入出力が共通しているという点で，Visual storytelling (Huang, Ferraro, Mostafazadeh, Ishan Misra, Devlin, Girshick, He, Kohli, Batra, Zitnick, Parikh, Vanderwende, Galley, and Mitchell 2016) に類似している．Visual storytelling では，図 1 と同様に写真列を入力としてシステムが各写真に対応する文章を出力する．この課題では写真から説明文を生成するキャプション生成 (You, Jin, Wang, Fang, and Luo

2016; Biten, Gomez, Rusinol, and Karatzas 2019) と違い，出力の文章は写真列の時系列を考慮した一貫性があることが要求される．本論文で取り扱う課題は Visual storytelling と比較して，出力のレシピは読者が読んで実行できるように，簡潔で具体的な記述であることが求められる．つまり，レシピにおける重要な物体や動作である食材，道具，調理者の動作を表す重要語と，それ含む表現が正しく生成されなければならない．例えば，図 1 の工程 1 においては，「ちくわ」や「切り」が重要語であるが，写真を説明するためには「1/3 の大きさに」といった表現も重要語に添えて生成する必要がある．これらをまとめて本論文では重要語を過不足なく含む表現と呼ぶ．

これらの重要語を過不足なく含む表現は，手順を記述する上で必要不可欠である．そのため，これらの表現は，手順に付与している写真の内容を大きく反映しているものと言える．この性質をもとに，料理ドメインでは完成写真に適したレシピを得る課題が検索課題として提案され，その解法として完成写真とレシピの間で共有された潜在的な意味に基づく特徴空間を学習する共有潜在空間モデルが高い性能を発揮してきた (Salvador, Hynes, Aytar, Marin, Ofli, Weber, and Torralba 2017; Zhu, Ngo, Chen, and Hao 2019; Chen and Ngo 2016)．しかしながら，完成写真とレシピの組ではなく，レシピの実行途中の写真と手順の組での共有潜在空間モデルは未だ提案されていない．この課題を解く場合，MSCOCO (Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár, and Zitnick 2014) や Flickr30k (Young, Lai, Hodosh, and Hockenmaier 2014) などの一般的なドメインにおける写真とその説明文を対象とする既存の共有潜在空間モデル (Wang, Li, Huang, and Lazebnik 2016) で写真と手順の組を用いて学習しても高い性能を得ることは難しい．これは次の手順で何を記述するか，またその際に特に言及する必要がある前の手順からの差分は何かといった文脈に大きく影響を受けるためであると考えられる．これらを考慮するために，写真に対応する手順だけでなく，レシピ全体を考慮できるように既存の共有潜在空間モデルの手順側のエンコーダに工夫を加える．この工夫によって，このモデルに写真を入力した時，近傍の手順には重要語を過不足なく含む表現の情報が含まれていると期待できる．これにより，各入力写真に対応する共有潜在空間上のベクトルは重要語を過不足なく含む表現が強調されたものとなることが期待できる．提案手法では，このような共有潜在空間を用いて写真の埋め込みベクトルを獲得した後，その空間中での近傍点を利用しながら文生成を行うことで，これらの表現を正しく生成する．

本手法を実装し，日本語のレシピを用いて評価実験を行なった．その結果，提案した共有潜在空間モデルは既存のモデルと比較して高い検索性能を得られた．また，レシピ生成の点においても，提案手法は BLEU，ROUGE-L，CIDEr-D といった生成文の自動評価尺度だけでなく，重要語を正しく生成できているかを測定した重要語生成の評価も Visual storytelling の標準的なベースラインを上回ることを実験的に確認した．そして，提案手法は写真に適した重要語を正しく生成していることを実例により確認した．考察では，提案手法が入力写真列に適したレシ

ピを生成することに成功したケースと失敗したケースを確認した．また，提案手法の重要な要素である，共有潜在空間についてのパラメータや，訓練データ量を変更した時の性能の変化を確認し，提案手法が性能を発揮する上で適当なパラメータやデータ量について検証した．

2 関連研究

入力と出力がそれぞれ写真列と文章であるという点において，本論文の課題設定は Visual storytelling (Huang et al. 2016) と類似している．Visual storytelling は，写真列を入力としてシステムが各写真に対応する文章を出力するという課題である．この課題では，システムは写真を説明する文章を生成するキャプション生成と異なり，写真列の前後関係を考慮した一貫性のある文章を出力することが求められる．Visual storytelling の課題に対して，Liu ら (Liu, Fu, Mei, and Chen 2017) は写真とテキストの共有潜在空間を学習しながら文生成する手法を提案している．この研究では共有潜在空間と文生成のモデルを同時学習しているのに対し，本研究では学習済みの共有潜在空間中に埋め込まれた手順ベクトルを直接文生成の時に参照している点が異なる．こうすることで，写真に適した重要語を過不足なく含む表現を有する手順ベクトルの情報を明示的に入力へ含めることができ，重要語を過不足なく含む表現を生成しながらレシピを生成することができる．

レシピの生成という課題としては，本研究での写真列を入力とする場合も含め，様々な研究がある．Salvador ら (Salvador, Drozdzal, Giro-i-Nieto, and Romero 2019) は完成写真からレシピを生成する手法を提案している．この研究では，材料予測器と文生成器を同時に学習させることによって，完成写真からレシピのタイトル，材料，レシピを全て生成することで，Web 上の完成写真から考えられるレシピの候補をユーザに提示するシステムを構築することを目的としている．Kiddon ら (Kiddon, Zettlemoyer, and Choi 2016) はレシピのタイトルと材料を入力として与え，生成文で材料を利用したかどうかを注意機構 (Luong, Pham, and Manning 2015) を用いて確認しながらレシピを生成する手法を提案している．しかし，本研究では材料だけでなく調理者の動作や道具も重要語として扱っている点が異なっており，これらを含めることで材料をどう扱うのかという点も考慮している．同じくタイトルと材料からレシピを生成する研究としては，Bosselut ら (Bosselut, Celikyilmaz, He, Gao, Huang, and Choi 2018) の研究が挙げられる．この研究では，参照文と生成文の類似度を報酬とし，強化学習を用いて破綻がないように長文 (レシピ) を生成する手法を提案している．概してこれらの研究では，レシピを高い精度で生成することよりも，いかに破綻させずに構造を持つ文書であるレシピを生成するかという点に着目している．そのため，入力に手順途中の情報が不足しており，十分な精度でレシピを生成するまでには至っていない．

一方で，Mori ら (Mori, Maeta, Sasada, Yoshino, Hashimoto, Funatomi, and Yamakata 2014a)

は手順の流れを重要語の有向グラフで表現したフローグラフ (Mori, Maeta, Yamakata, and Sasada 2014b) を入力としてレシピを生成する手法を提案している．手順途中の情報が与えられていない既存研究と比較し，実用的な精度でレシピを生成することに成功している．フローグラフではなく写真列を中間情報として与えるメリットとして，レシピを得ることが調理者にとって容易である点が挙げられる．フローグラフからレシピを得る場合，フローグラフをアノテーションし用意する必要がある．一方，写真列を入力とする場合は，調理者はレシピを実行する上で重要な写真を撮るだけでレシピを得ることができる．そのため，本研究では手順途中の状態を考慮するために，写真列を入力として与えている．同じく写真列である料理動画 (フレーム列) を入力としてレシピを得る研究として，Ushiku ら (Ushiku, Hashimoto, Hashimoto, and Mori 2017) の研究がある．この研究と本研究との違いは，扱っている写真列が異なっている点である．Ushiku ら (Ushiku et al. 2017) の研究で入力とする料理動画は，キッチン全体を撮影した未編集の料理映像であり，手順実施の上で直接関係のないフレームや，調理者が待機しているフレームなどが多く含まれるため，生成すべき手順数は未知でかつ重要なフレームを予測しながら文生成を行う必要性がある．そのため，この重要フレームの予測誤差や文生成自体のミスなどの影響により，生成した手順書の精度は実用的な精度には至っていない．一方，本研究で対象とする写真列は，調理者が少なくとも手順実施の上で重要な場面で写真を撮影しており，これらの写真に 1 つの手順が紐づいているため，生成すべき手順数が既知である．これらの観点から，本論文で入力とする写真列には手順途中の情報が十分に含まれており，こうして撮影された写真に適した手順を生成することで，より高い精度でレシピを生成することが期待できる．本研究と同じく料理ドメインで写真列を入力として取り扱った研究として，Yagcioglu ら (Yagcioglu, Erdem, Erdem, and Ikizler-Cinbis 2018) の RecipeQA がある．この研究では，コンピュータが写真付きレシピをどの程度理解できるかを測定するためのマルチメディア QA データセットを提案している．Chandu ら (Chandu, Nyberg, and Black 2019) は本研究と同じく写真列を入力としてレシピを生成する研究を行なっている．この研究では，有限オートマトンモデルを用いて一貫性のあるレシピを生成することに着目したのに対し，本研究では，重要語を過不足なく含む表現が生成できているかという点に着目しレシピを生成している点が異なる．本研究ではレシピを実施する上で重要な単語である食材，道具，調理者の動作を重要語と定義し，これを含む表現を生成する手法を提案する．このように，写真を説明する上で重要な単語を生成することに焦点を当てた研究はキャプション生成の分野において行われている．Biten ら (Biten et al. 2019) は，写真からテンプレートを生成したのち，ニュース記事から人名や場所といった固有表現をテンプレートに当てはめて重要な単語を生成する手法を提案している．また，You ら (You et al. 2016) の研究では，写真から重要な単語を多クラス分類などを用いて抽出し，それらの単語に注意機構 (Luong et al. 2015) を用いながらキャプション生成を行う手法を提案している．本研究ではこの重要な単語を含む表現の生成に学習済みの共有潜在空間を用いてい

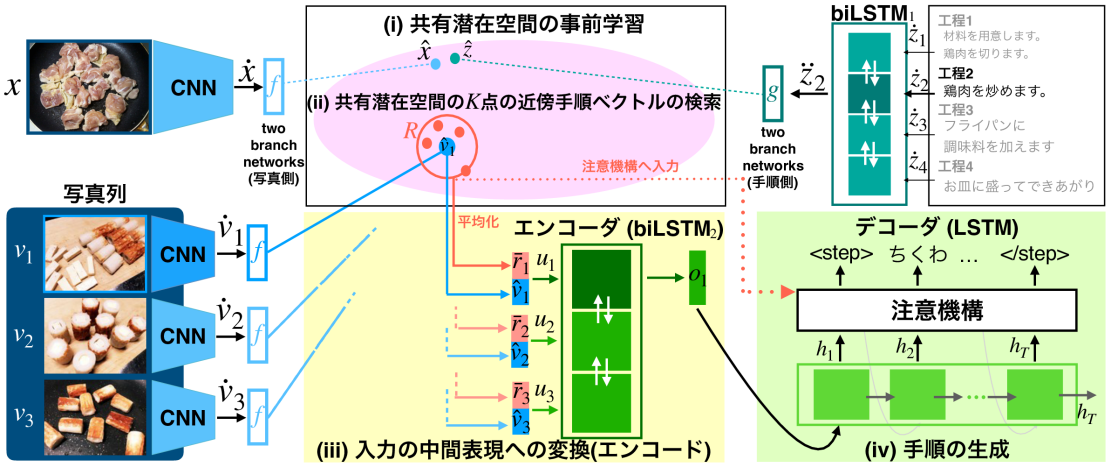

図 2　提案手法の概要.

る．この共有潜在空間上の各写真のベクトル，各写真に類似した手順には重要語を過不足なく含む表現の情報を含んでいると考えられ，これらを用いることで重要語を過不足なく含む表現を生成する手法を提案する．

3　提案手法

本章では，写真列を入力として，写真列に適した重要語を過不足なく含む表現を持つレシピを生成する手法について説明する．手順を記述する上で，これらの表現は欠かせない要素である．そのため，これらの表現は，手順に付与した写真の内容を大きく反映している．この性質をもとに，本論文では検索課題として取り組まれてきた手法を文生成の手法へ組み込むことで，重要語を過不足なく含む表現を持つ情報を利用しながらレシピを生成する手法を提案する．図 2 に提案手法の概要を示す．提案手法は以下の 4 つのプロセスで構成されている．

(i)　Web 上に存在する大量の写真と手順の組を用いて共有潜在空間をあらかじめ学習させておく．この時，前後の手順の文脈を考慮するために，共有潜在空間の手順側のエンコーダへ biLSTM を追加し学習する．その後，入力された写真列の各写真ごとに以下の (ii) から (iv) を繰り返して手順を 1 つずつ生成し，生成した全ての手順をあわせてレシピとして出力する．

(ii)　写真に適した重要語を過不足なく含む表現を有する手順ベクトルを得るために，共有潜在空間上における入力写真のベクトルを得た後，そのベクトルをもとに近傍の K 個の手順ベクトルを検索する．

(iii) 検索した K 個の手順ベクトルを平均したベクトルと写真の埋め込みベクトルを結合し，手順間の時系列を考慮したベクトルを biLSTM を用いて計算する．

(iv) 最後に，写真ごとに手順を出力する．

3.1 共有潜在空間

最初に，(i) の学習を行うために，Wang ら (Wang et al. 2016) によって提案された Two branch networks を利用する．このモデルは，写真側，テキスト側にそれぞれ非線形な活性化関数と 2 層の多層パーセプトロンからなるニューラルネットワークを用いて，写真，テキストの間で共有された潜在的な意味に基づく特徴空間を学習する．この特徴空間では，写真に対してその写真の内容を記述する手順は近く位置する．そうでない場合，遠く位置することとなる．そのため，このモデルに手順と写真を与えることで，手順と写真の間での類似度を計算することができる．一般のテキストと異なり，レシピ中では，非常に多くの物体名がゼロ照応の形で省略される (Malmaud, Wagner, Chang, and Murphy 2014)．Malmaud らが分析対象としたレシピは英語のレシピであるが，本研究で対象とする日本語のレシピの中でも同様の傾向が見られる．例えば，図 1 の工程 3 において，“きつね色になるまで焼きます” という手順中では，焼く対象である「チーズを挟んだちくわ」が省略されている．これらの省略のため，本研究の予備実験では元の Two branch networks では高い性能を得ることができなかった．この問題を解決するために，biLSTM を手順側に挿入することで，写真に対応する手順だけではなく，前の手順全体と後の手順全体も考慮することができるように変更を加える．これにより，省略された物体や，代名詞で表現される物体名を前後の手順から情報として加えることができるようになり，性能悪化を防ぐことができるようになる．入力の写真を $\boldsymbol{x}$，M 個の手順からなる手順列を $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m, \ldots, \boldsymbol{z}_M)$ と置き，この変更を以下のように数式を用いて表現する．写真 $\boldsymbol{x}$ を畳み込みニューラルネットワーク (CNN) に入力して得られた特徴ベクトルを $\dot{\boldsymbol{x}}$，各手順中の各単語をあらかじめ訓練データのレシピ全体で学習しておいた word2vec (Mikolov, Sutskever, Chen, Corrado, and Dean 2013) で分散表現に変換し，その平均ベクトルを手順の特徴ベクトル $\dot{\boldsymbol{z}}_m$ とする．Two branch networks の写真側，手順側のニューラルネットワークをそれぞれ f, g と置くと，写真，手順の共通潜在空間での埋め込みベクトル $\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}$ は以下のように表される．

$$\dot{\boldsymbol{x}} = \mathrm{CNN}(\boldsymbol{x}) \tag{1}$$

$$\hat{\boldsymbol{x}} = f(\dot{\boldsymbol{x}}) \tag{2}$$

$$\ddot{\boldsymbol{z}}_m = \mathrm{biLSTM}_1(\dot{\boldsymbol{z}}_1, \dot{\boldsymbol{z}}_2, \ldots, \dot{\boldsymbol{z}}_m, \ldots, \dot{\boldsymbol{z}}_M) \tag{3}$$

$$\hat{\boldsymbol{z}} = g(\ddot{\boldsymbol{z}}_m) \tag{4}$$

ここで, CNN(·) は図中の CNN に対応し, biLSTM$_1$(·) は追加した biLSTM である図中の biLSTM$_1$ に対応する．また, $\tilde{\boldsymbol{z}}_m$ は biLSTM$_1$ の m 番目の出力ベクトルを表す．得られたベクトル $\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}$ を用いて，Two branch networks の損失関数として提案されている，構造を保った Triplet margin loss (Balntas, Riba, Ponsa, and Mikolajczyk 2016) を損失関数として最適化するように biLSTM$_1$，f，g，CNN の重みを更新する．この時，word2vec の重みのみ固定して学習する．また，Two branch networks と同様に距離関数として余弦距離を用いた．学習後，以下のレシピ生成においてはこれらの重みを固定して利用する．

3.2 レシピ生成

入力の写真列を $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n, \ldots, \boldsymbol{v}_N)$，出力の手順列を $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n, \ldots, \boldsymbol{y}_N)$ とする．n 番目の写真 $\boldsymbol{v}_n$ を CNN へ入力して特徴ベクトル $\dot{\boldsymbol{v}}_n$ へ変換し，さらにそれを Two branch networks の写真側のニューラルネットワークに入力する．こうすることで，共有潜在空間上の写真の埋め込みベクトルを得ることができる．共有潜在空間上の写真の埋め込みベクトル $\hat{\boldsymbol{v}}_n$ は対応する手順との距離が近くなるように学習されているため，重要語を過不足なく含む表現の情報を考慮したベクトルを得ることが期待できる．この処理は CNN(·), f を用いて以下のように表される．

$$\dot{\boldsymbol{v}}_n = \mathrm{CNN}(\boldsymbol{v}_n) \tag{5}$$

$$\hat{\boldsymbol{v}}_n = f(\dot{\boldsymbol{v}}_n) \tag{6}$$

得られた写真列の各埋め込みベクトルを用いて，(ii) から (iv) の手続きで手順を生成する．以下に，それぞれのプロセスの詳細を述べる．

(ii) 共有潜在空間上の手順ベクトルの検索: 写真の埋め込みベクトル $\hat{\boldsymbol{v}}_n$ をもとに，その近傍の手順ベクトルを K 個，共有潜在空間の学習に利用したデータセットから検索する．得られた K 個のベクトルを，$R = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_K)$ とする．手順ベクトルの平均ベクトル $\bar{\boldsymbol{r}}_n$ は以下のように計算される．

$$\bar{\boldsymbol{r}}_n = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{r}_k \tag{7}$$

最後に，得られた手順ベクトルの平均ベクトルと，写真の埋め込みベクトルを結合する．こうして得られるベクトルを，$\boldsymbol{u}_n = (\hat{\boldsymbol{v}}_n, \bar{\boldsymbol{r}}_n)$ と書く．

(iii) 入力の中間表現への変換: 写真列の時系列の情報を考慮するために，(ii) で各写真ごとに得られたベクトル $\boldsymbol{u}_n$ を写真列のエンコーダに入力する．ここで，前の手順だけでなく後ろの手順も考慮するために，エンコーダには biLSTM を用いる．

$$\boldsymbol{o}_n = \mathrm{biLSTM}_2(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n, \ldots, \boldsymbol{u}_N) \tag{8}$$

ここで，$\mathrm{biLSTM}_2(\cdot)$ は写真列のエンコーダの biLSTM である，図中の biLSTM_2 に対応する．

(iv) 手順の生成: LSTM をデコーダとして用いる．(iii) で得られた $\boldsymbol{o}_n$ を入力として，手順の開始記号 ($\langle\mathrm{step}\rangle$) から終端記号 ($\langle/\mathrm{step}\rangle$) が生成されるまで，単語を一つ一つ出力し，手順を生成する．単語を出力する際に，検索した K 個の手順ベクトルを参照しながら単語を選択するために，Luong らの注意機構 (Luong et al. 2015) をモデルに組み込む．注意機構を用いることで，各手順ベクトルから必要な情報を参照しながら単語を選択できるため，より重要語を過不足なく含む表現を生成しやすくなると期待できる．手順ベクトル $\boldsymbol{r}_k$ と，デコーダの隠れ層 $\boldsymbol{h}$ の間での注意機構の重みを計算するために，Luong らの注意機構の中から general attention を用いる．手順の t 番目の単語を出力する時の隠れ層のベクトル $\boldsymbol{h}_t$ と，検索された手順ベクトル R から計算される k 個目の手順ベクトルへの注意機構の重み a_t^k，文脈ベクトル $\boldsymbol{c}_t$，それらから得られる注意ベクトル $\tilde{\boldsymbol{h}}_t$ は，以下のように書くことができる．

$$a_t^k = \frac{\exp\left(\boldsymbol{r}_k^T \boldsymbol{W}_a \boldsymbol{h}_t\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{r}_j^T \boldsymbol{W}_a \boldsymbol{h}_t\right)} \tag{9}$$

$$\boldsymbol{c}_t = \sum_{k=1}^{K} a_t^k \boldsymbol{r}_k \tag{10}$$

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_c(\boldsymbol{c}_t, \boldsymbol{h}_t)) \tag{11}$$

ここで，$\boldsymbol{W}_a$ と $\boldsymbol{W}_c$ は学習によって得られる重み行列である．これらの式より，出力単語の条件付き確率分布 $p(y_t|y_{<t}, \boldsymbol{o}_n)$ はソフトマックス関数を用いて以下のように書くことができる．

$$p(y_t^n|y_{<t}^n, \boldsymbol{o}_n) = \mathrm{softmax}(\boldsymbol{W}_o \tilde{\boldsymbol{h}}_t + \boldsymbol{b}_o) \tag{12}$$

ここで，$\boldsymbol{W}_o$ は注意ベクトル $\tilde{\boldsymbol{h}}_t$ を語彙サイズのベクトルへ変換する重み行列であり，$\boldsymbol{b}_o$ はバイアスを表す．推論する際には，条件付き確率分布の中で最も確率が高い単語を語彙から選択し，出力する．また，1 つの手順を出力した後，デコーダの最後の隠れ層は次の手順を生成する時の最初の隠れ層として設定される．

損失関数の計算: 学習を行うときは，写真列 $\boldsymbol{V}$ と手順列 $\boldsymbol{Y}$ の対の集合である訓練データ $\mathcal{D}$ に対して，以下の負の対数尤度の合計が最小になるように学習を行う．

$$L(\boldsymbol{\theta}) = -\sum_{(\boldsymbol{V},\boldsymbol{Y})\in\mathcal{D}} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n|\boldsymbol{v}_n;\boldsymbol{\theta}) \tag{13}$$

ここで，$\boldsymbol{\theta}$ は biLSTM_2，$\boldsymbol{W}_a$，$\boldsymbol{W}_c$，$\boldsymbol{W}_o$，$\boldsymbol{b}_o$，LSTM の重みを表す．

	訓練	検証	評価
レシピ数	183,502	20,340	22,702
平均工程数/レシピ	5.93	5.85	5.94
平均単語数/レシピ	134.72	133.30	134.50
平均文数/手順	1.50	1.51	1.50
最大文数/手順	11	9	10
最小文数/手順	0	0	0
語彙サイズ		26,285	

表 1 データセットの統計結果．手順あたりの文数は GiNZA (`https://megagonlabs.github.io/ginza/`) を用いて分割を行った結果をもとに数えている．なお，文数 0 は手順が空白のみの場合を表す．

4 評価

4.1 データセット

実験で利用するデータセットは，Cookpad Image Dataset (Harashima, Someya, and Kikuta 2017) である．Cookpad Image Dataset は，Cookpad のユーザによって投稿された日本語のレシピと，レシピの完成写真，また，レシピの各手順に対して付与された写真からなるデータセットである．このデータセットから全ての手順に写真がアップロードされているものだけを対象に抽出した [1]．その結果，約 20 万のレシピを得ることができた [2]．得られたレシピを学習:検証:評価に 8:1:1 で分割し，以下のようにデータセット中の全ての写真，レシピそれぞれに前処理を行い，共有潜在空間とレシピ生成を同じ分割単位で学習し評価した．写真の前処理として，CNN に入力するために写真の縦と横の長さで大きい方が 256 になるようにアスペクト比を保ったままサイズを変換した後，写真の中央からサイズが 224×224 になるように切り抜いた．レシピの前処理として，KyTea (Neubig, Nakata, and Mori 2011) を用いて単語分割を行った．この分割結果に対し，訓練データで出現頻度数が 3 回以下の単語は未知語とした．以下のレシピの生成においては，KyTea によって得られた単語の分割単位で生成を行う．表 1 にデータセットの統計情報を示す．

4.2 詳細設定

共有潜在空間への写真側のエンコーダとして，ImageNet (Deng, Dong, Socher, Li, Li, and Fei-Fei 2009) で学習済みの ResNet-50 (He, Zhang, Ren, and Sun 2016) を用いた．ResNet-50 の最終層のソフトマックス層を取り除いたため，写真側の出力の次元数は 2,048 である．共有潜在

[1] Cookpad Image Dataset は約 310 万の写真と手順からなるデータセットである．しかし，手順の中には写真がアップロードされていないものもある．

[2] データセットの Cookpad Image Dataset との対応は `http://www.ar.media.kyoto-u.ac.jp/member/nishimura/recipe_ids.html` からダウンロードできる．

	画像から手順の検索				手順から画像の検索			
	MedR	R @ 1	R @ 5	R @ 10	MedR	R @ 1	R @ 5	R @ 10
ランダムに選択	500	0.001	0.005	0.010	500	0.001	0.005	0.010
biLSTM_1 なし (Wang et al. 2016)	21	0.088	0.259	0.375	21	0.079	0.254	0.358
biLSTM_1 あり	**7**	**0.170**	**0.445**	**0.573**	**7**	**0.167**	**0.438**	**0.581**

表 2　two branch networks へ biLSTM_1 を追加したときの共有潜在空間の検索性能の変化.

空間でのテキスト側のエンコーダの biLSTM_1 の隠れ層の次元数は 1,024 としたため，出力の次元数は双方向の出力ベクトルを結合し，2,048 次元となる．学習手順は Two branch networks と同じく，バッチサイズは 1,500 にし，ネガティブサンプリングの際にはミニバッチ内で最も損失関数の値が大きい 50 サンプルのみから損失関数を計算し，共有潜在空間の学習を行った (Wang et al. 2016)．文生成のモデルでは，隠れ層の次元数をエンコーダとデコーダ共に 512 に設定した．学習時には，共有潜在空間の重みは固定し，その他の重みは Adam (Kingma and Ba 2015) を用いて最適化を行なった．なお，バッチサイズは 64 とし，Adam の初期値は $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.99$ として設定した．毎エポックの終わりに検証用データセットで負の対数尤度を計算し，3 エポック連続で負の対数尤度が下がらなかった場合に学習を停止した．また，検索する入力写真の近傍の手順ベクトル数は $K = 10$ とした．これらのハイパーパラメータは，検証用データセットを用いて決定した.

4.3　結果

提案手法を評価するために，以下の 4 点を評価した.

(i) **共有潜在空間への biLSTM_1 の影響**: 前後の手順を考慮することによる共有潜在空間の性能の変化を検索課題として評価した.

(ii) **生成文の自動評価尺度による評価**: 生成したレシピと正解のレシピの単語レベルの一致率を BLEU，ROUGE-L，CIDEr-D を用いて評価した.

(iii) **重要語の性能評価**: 生成したレシピは読者が読んで実行できるよう，前後の手順との差分を考慮しながら写真に沿った重要語を過不足なく含む表現を有する手順を生成しなければならない．そのため，参照文中の重要語をモデルが正しく言及できたかどうかを評価した.

(iv) **定性的評価**: 実際に写真列から生成したレシピの一例を示し，提案手法の有用性を確認した.

4.3.1　共有潜在空間への biLSTM_1 の影響

最初に，Two branch networks へ biLSTM_1 を追加したことによる影響を以下のように評価して確認した．まず，Two branch networks や im2recipe (Salvador et al. 2017) といった複数の従来手法が慣例的に評価データの数を 1,000 程度としているため，これに習う形で，ランダムに

		BLEU1	BLEU4	ROUGE-L	CIDEr-D
ベースライン	写真列	17.8	3.2	17.1	12.4
	写真列 + タイトル	25.1	4.4	17.2	15.7
	写真列 + タイトル + 材料	27.4	5.6	20.3	16.2
提案手法 biLSTM$_1$(なし)	写真の埋め込みベクトル	22.2	4.1	20.1	13.2
	写真の埋め込みベクトル + Top1 の手順ベクトル	30.9	6.3	21.6	17.8
	写真の埋め込みベクトル + Top$K$ の手順ベクトル	26.2	5.4	21.0	17.4
提案手法 biLSTM$_1$(あり)	写真の埋め込みベクトル	26.8	5.7	21.2	16.0
	写真の埋め込みベクトル + Top1 の手順ベクトル	28.7	6.2	22.2	19.3
	写真の埋め込みベクトル + Top$K$ の手順ベクトル	**31.3**	**6.6**	**22.4**	**20.3**

表 3　自動評価尺度による生成したレシピの評価結果．実験では，$K = 10$ として設定した．

テストセットから 1,000 個の写真と対応する手順のペアを取り出した．次に，写真を入力とした時には，手順集合を余弦類似度の降順にソートし，写真のペアとなる手順が現れる順位の中央値 (MedR) と，写真のペアとなる手順が上位 k 番目以内に現れる割合 (Recall@k) を計算した．なお，手順を入力にした際にも，同様の基準で評価した．この結果を表 2 に示す．この結果から，biLSTM$_1$ がないオリジナルの Two branch networks に比べ，大きく性能が向上していることがわかる．よって，biLSTM$_1$ を追加したことによってモデルが手順間の文脈を参照することができるようになったため，性能を改善することができたと考えられる．

4.3.2　定量的評価

提案手法を評価するために，評価データの全てのレシピを用いて文生成の自動評価尺度である BLEU1，BLEU4，ROUGE-L，CIDEr-D を評価した．この時，生成したレシピと正解のレシピの間で手順同士で評価したのではなく，レシピ同士で評価した．また，評価に用いる単語の分割単位は KyTea で処理して得られた単語の分割単位で評価した．検索課題として取り組まれてきた手法を文生成の手法に組み込んだことによる性能の変化を見るため，Visual storytelling (Huang et al. 2016) で挙げられている，写真列をエンコーダの biLSTM へ入力し，デコーダの LSTM で出力するニューラルネットワークをベースラインとした．加えて，レシピに付与した材料やタイトルの各単語を word2vec (Mikolov et al. 2013) を用いて分散表現に変換し，その平均ベクトルを写真列の ResNet-50 の出力ベクトルと結合し，エンコーダへの入力に加えたベースラインも用意した (表 3 中の「写真列+タイトル」および「写真列+タイトル+材料」)．なお，表 3 中の提案手法の「写真の埋め込みベクトル」とベースラインの「写真列」の違いは共有潜在空間を利用するか否かが異なる．前者では，共有潜在空間中の写真のベクトルを用いているのに対し，後者では ResNet-50 の出力ベクトルをそのまま利用している．また，提案手法においてはモデルが入力写真をもとに手順ベクトルの検索を行うが，この検索先となる手順ベクトルは訓練データのものとした．表 3 に評価結果を示す．この結果により，提案手法が全ての指標でベースラインと比較して性能が向上したことを確認した．

		F	T	Ac	合計
ベースライン	再現率	8.7	18.0	16.5	13.4
	適合率	12.8	20.0	27.9	20.3
	F 値	10.4	18.9	20.8	16.2
提案手法 (Top1) biLSTM$_1$(あり)	再現率	23.1	30.7	20.3	22.6
	適合率	22.4	20.5	22.7	22.3
	F 値	22.7	24.6	21.5	22.4
提案手法 (Top$K$) biLSTM$_1$(あり)	再現率	49.1	33.3	31.0	38.7
	適合率	41.6	26.0	28.6	33.7
	F 値	**45.1**	**29.3**	**29.7**	**36.0**

表 4　重要語生成の性能評価．表中のベースラインは，表 3 中の「写真列+タイトル+材料」を示す．

4.3.3　重要語生成の性能評価

前節で測定した BLEU，ROUGE-L，CIDEr-D のような自動評価尺度に加えて，正解のレシピの重要語が，生成したレシピ中に正確に現れる割合を評価し，提案手法が写真に適した重要語を生成しているかどうかを評価した．この時，正解のレシピを比較対象としている．レシピにおいて物体名を省略することが頻繁に行われる理由として，手順の説明においては，前の状態からの差分にしか言及しないためであると考えられ，同様の理由から，前の状態との差分として言及すべき重要語が決定する．そのため，実際のレシピで言及されている重要語は状態の差分を反映しており，これが過不足なく言及されているかどうかを通して，レシピ文としての質を評価できる．料理ドメインにおいては，レシピフローグラフコーパス (Mori et al. 2014b) によると，食材 (F)，道具 (T)，調理者の動作 (Ac) の 3 つがレシピ中に統計的に多く出現することが確認されている．よって，これらのカテゴリに属す単語を重要語とし，生成したレシピと正解のレシピに現れる重要語リストから以下のように計算される適合率，再現率，F 値を測定することで，重要語生成の性能を評価した．

$$\text{再現率} = \frac{\text{正解の重要語数}}{\text{参照文中のレシピに現れる重要語数}} \tag{14}$$

$$\text{適合率} = \frac{\text{正解の重要語数}}{\text{生成したレシピに現れる重要語数}} \tag{15}$$

$$F\text{ 値} = \left(\frac{\text{再現率}^{-1} + \text{適合率}^{-1}}{2}\right)^{-1} \tag{16}$$

しかしながら，この評価は同義語や表記揺れの問題から，自動的に計算することができない．そのため，レシピを 50 個ランダムで評価データから抽出し，手動で同義語と表記揺れのみを修正し，笹田らの研究で定義されている基準 (笹田，森，山肩，前田，河原 2015) を参考に単語中

から F，T，Ac に限ってタグを割り当て，重要語生成の性能を評価した [3]．表 4 にその結果を示す．なお，表中のベースラインは表 3 中の「写真列+タイトル+材料」を，Top1 は提案手法の項目の，biLSTM$_1$(あり) の「写真への埋め込みベクトル+Top1 の手順ベクトル」を，Top$K$ は，提案手法の項目の，biLSTM$_1$(あり) の「写真への埋め込みベクトル+Top$K$ の手順ベクトル」を表す．また，検索する手順ベクトル数 K は 10 である．この表より，Top1 の手法は明確にベースラインの結果を上回り，また Top$K$ はさらにその Top1 を上回るという結果となった．よって，提案手法が重要語を正しく生成しながらレシピを生成していると言える．

4.3.4 定性的評価

図 3 に入力の写真列と，ベースライン，提案手法によって生成されたレシピ，そして正解のレシピを載せる．この図より，ベースラインは写真に適したレシピを生成することに失敗している一方で，提案手法は写真に適した重要語を過不足なく含む表現を生成しているということが分かる．

4.4 考察

次に，提案手法によって生成されたレシピと，正解のレシピを見比べることによって提案手法が手順を生成することに成功したケースと，生成に失敗したケースをそれぞれ取り上げ，提案手法の長所と短所を定性的に考察を行う．また，提案手法の性能に関わる重要な要素である，共有潜在空間へ検索する手順ベクトルの数 K や，訓練データ量を変更することによる性能の変化を確認し，提案手法を実現するために必要な手順ベクトル数や訓練データ量を検証する．

4.4.1 重要語を過不足なく含む表現の生成に成功したケース

注意機構を用いることで，モデルは各単語を出力するときに共有潜在空間から検索した K 個の手順ベクトルの中から，どの手順を参照して単語を選択するのかを各手順ベクトルに重みを付けて表現することができる．図 4 に重要語を過不足なく含む表現の生成に成功した手順，K 個の入力写真の近傍手順ベクトル，注意機構による手順ベクトルへの重みを可視化したものを示す．提案手法によって重要語を過不足なく含む表現の生成に成功するケースでは，重要語である「たまねぎ (F)」や，「切り (Ac)」，加えて切り方を示す「みじん」と言った重要語を過不足なく含む表現を有する手順ベクトルに対して高い重みが割り当てられている．一方で，検索した手順ベクトル中の，重要語を過不足なく含む表現を生成する上で不要である手順については，低い重みが割り当てられていることが分かる．このことから，注意機構を導入したことに

[3] 重要語のアノテーションの結果，アノテーションした重要語の数は合計でそれぞれ F は 288，T は 85，Ac は 347 得られた．これらのアノテーション結果は http://www.ar.media.kyoto-u.ac.jp/member/nishimura/annotation.html からダウンロードできる．

図 3　レシピの生成例．太字かつ二重線で書かれた箇所は正しく重要語が生成された箇所であり，下線部付きで書かれた箇所は重要語生成に失敗した箇所である．また，重要語のタグの種類 (F (食材)，T (道具)，Ac (調理者の動作)) を左下に付与した．

よって，モデルは手順ベクトル中から必要な情報を参照しながら重要語を過不足なく含む表現を生成していることが分かる．

4.4.2　重要語を過不足なく含む表現の生成に失敗したケース

図 3 の結果にもあるように，正しく手順を生成できなかったケースも存在する．図 5 に示すような例では，入力した写真で検索して得られた近傍の手順ベクトルに含まれる重要語は「薄

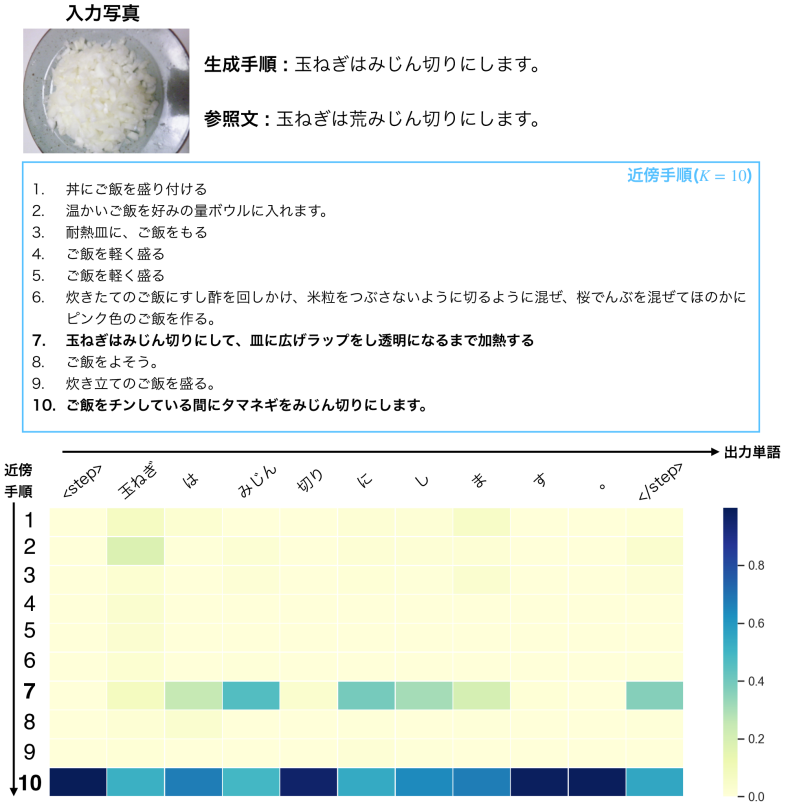

図 4　提案手法が重要語を過不足なく含む表現の生成に成功したケース．重要語である「玉ねぎ (F)」や「切り (Ac)」が現れる手順 (図中の手順 7，手順 10) に高い重みが割り当てられ生成されている．

揚げ (F)」ではなく，「パン (F)」として得られてしまったため，重要語を過不足なく含む表現を生成することができなかった．このように，入力した写真に近傍の手順ベクトル中に重要語を過不足なく含む表現が存在しない場合は，モデルが手順を正しく生成することは難しいと考えられる．

4.4.3　検索手順ベクトル数 K を変更した時の性能の変化

検索する手順ベクトル数 K を変更して学習し，評価することによって，必要な手順ベクトル数 K を考察する．表 5 に検索手順ベクトル数 K を変更した時の，生成文の自動評価尺度の変化を示す．なお，表中の $K > 1$ の場合は注意機構がモデルに組み込まれているが，$K = 1$ の場合，注意機構はモデルに組み込まれていない．この実験の結果，$K = 25$ の場合に BLEU1，BLEU4，ROUGE-L で最も高い性能を示し，$K = 10$ の場合に CIDEr-D で最も高い性能を示した．注意機構がモデルに組み込まれている時，K が 5 から 25 までの間は，K を大きくするにつれて性

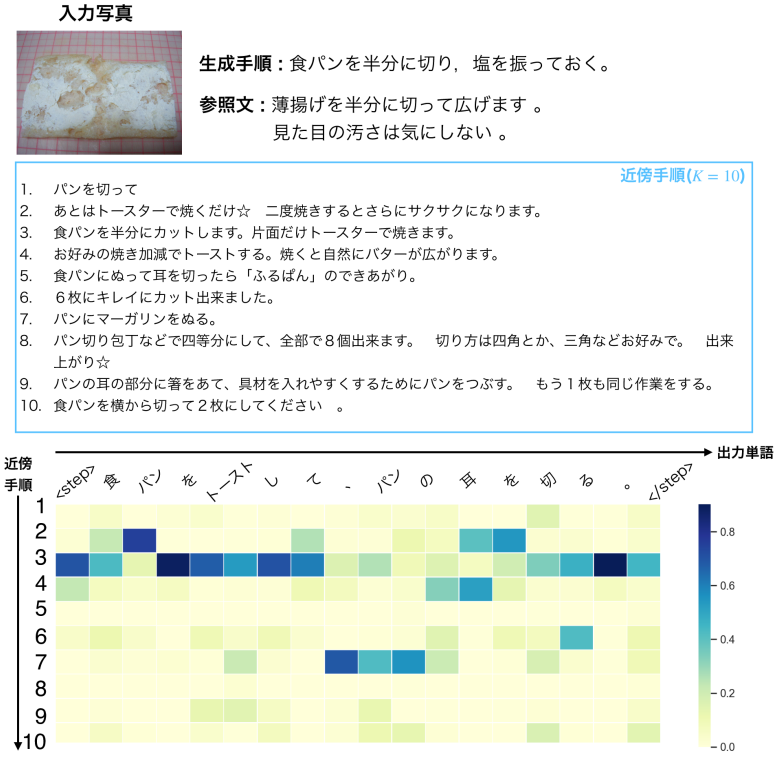

図 5　提案手法が重要語を過不足なく含む表現の生成に失敗したケース．入力画像で検索した時に得られた手順ベクトルが「食パン (F)」と得られ，正しく「薄揚げ (F)」を得られなかった．

		BLEU1	BLEU4	ROUGE-L	CIDEr-D
注意機構なし	K=1	28.7	6.2	22.2	19.3
注意機構あり	K=5	29.2	6.3	22.0	19.7
	K=10	31.2	6.6	22.4	**20.3**
	K=25	**31.3**	**6.8**	**23.1**	19.7
	K=50	29.6	6.6	22.8	19.6
	K=100	31.0	6.7	22.7	18.2

表 5　検索した手順ベクトル数 K を変化させた時の自動評価尺度の結果．

能が全体的に上昇する傾向にあることが分かる．一方で，K が 25 を超えた $K = 50$，$K = 100$ のモデルにおいては，$K = 25$ のモデルと比べて低い性能を示している．このことから，K の数を小さく設定すると，写真に適した重要語を過不足なく含む表現を有する手順ベクトルの検索に失敗し，性能が低くなるが，一方で，K の数を大きく設定すると，写真に適した重要語を過不足なく含む表現を有する手順ベクトルの他に写真に適さない手順ベクトルも手順生成に用

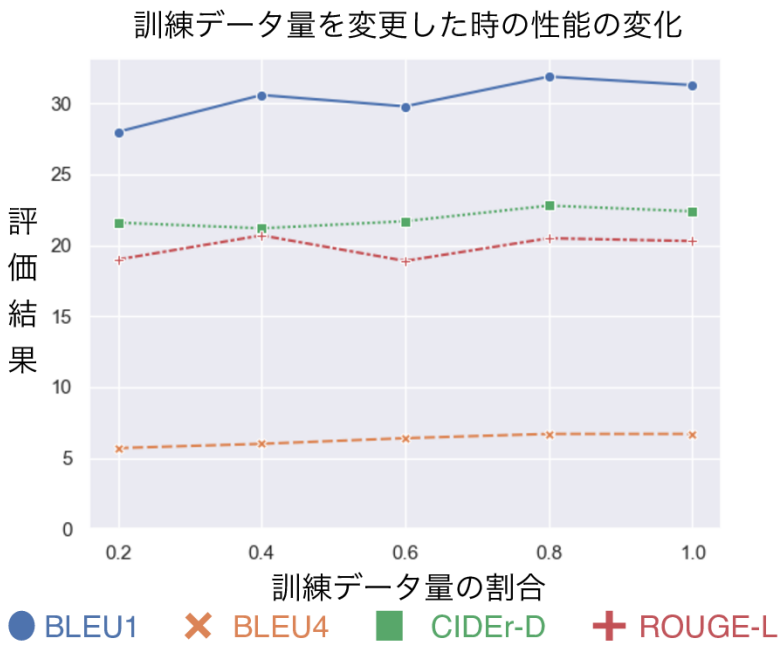

図 6　訓練データ量を変更した時の性能の変化.

いることとなる．そのため，モデルの性能が低下したものと考えられる．

4.4.4　訓練データ量を変更した時の性能の変化

モデルが入力写真をもとに手順ベクトルを検索する時，訓練データの中から検索を行う．そのため，学習に利用した訓練データ量は，モデルの性能を決める重要な要素であると言える．よって以下では訓練データ量を変更して学習し評価することによって，モデルが性能を発揮する上で必要なデータ量を考察する．図 6 に，訓練データ量を 20%，40%，60%，80%，100%の割合で変更した時の，生成文の自動評価尺度の変化を示す．また，この実験においては検索する手順ベクトル数 $K = 10$ として実験を行なった．この実験の結果，以下の点が明らかになった．

(i)　BLEU4 においては，ほぼ横ばいであるが訓練データ量の割合が 20%から 80%にかけて上昇し，100%で 80%の結果と横ばいとなった．

(ii)　BLEU1，ROUGE-L，CIDEr-D においては，訓練データ量の割合が 80%の時に BLEU1，ROUGE-L が最も高い性能を示した． また，CIDEr-D に関しては 40%の時に最も高い性能を示した．

以上の結果より，80%の訓練データ量で学習した時，BLEU1，BLEU4，ROUGE-L で最も高い性能を示した．CIDEr-D においては 40%の時に最も高い性能を示したが，80%の時との差は 0.2

ポイント (80%: 20.5，40%: 20.7) であり，差は小さい．よって，全体のデータ量の内 80%以上の訓練データ量を用いることで提案手法の性能を発揮することができると考えられる．

5 まとめ

本論文では，写真付きレシピの作成を支援するために，写真列からレシピを生成する課題を提案する．この課題では，生成したレシピは読者が読んで実行できるように重要語を過不足なく含む表現を正しく言及されていなければならない．これを達成するために，本論文では共有潜在空間モデルを文生成モデルに組み合わせる手法を提案する．従来の共有潜在空間モデルは一般的なドメインにおける写真と説明文を対象としていた．しかし，レシピにおいて，前後の手順の文脈に応じて言及すべき語が決まるため，前後の手順を考慮せずに写真と手順を 1:1 で学習する既存手法では高い性能を発揮できなかった．これを解決するために，前後の手順を考慮できるように既存の共有潜在空間モデルへ工夫を加える．こうして得られたモデルに写真を入力した時，近傍の手順は重要語を過不足なく含む表現の情報を有すると期待でき，また各入力写真に対応する共有潜在空間上の埋め込みベクトルは重要語を過不足なく含む表現が強調されたものとなることが期待できる．よって，この写真の埋め込みベクトルと，その空間中での近傍点を利用しながら文生成を行うことで，重要語を過不足なく含む表現を生成する手法を提案する．本手法を実装し，日本語のレシピを対象に評価実験を行なった．その結果，提案した共有潜在空間モデルは既存のモデルと比較して高い検索性能を得られた．また，レシピ生成の点においても，提案手法は BLEU，ROUGE-L，CIDEr-D といった生成文の自動評価尺度だけでなく，重要語を正しく生成できているかを測定した重要語生成の評価も Visual storytelling の標準的なベースラインを上回ることを実験的に確認した．そして，提案手法は写真に適した重要語を過不足なく含む表現を正しく生成していることを実例により確認した．考察では，重要語生成に成功したケースと失敗したケースを見比べることで，提案手法の長所と短所を明らかにした．さらに，提案手法が性能を発揮する上で必要な要素である，検索する手順ベクトル数 K や訓練データ量を変更した時の性能の変化を示し，必要な手順ベクトル数や訓練データ量を検証した．以上の実験と考察の結果，本論文の提案手法を用いて写真列に適したレシピを得ることができていることを実験的に示した．

謝辞

本研究はクックパッド株式会社の協力の下に行われたものである．よってここに感謝の意を表する．本論文の内容の一部は，The 12th International Conference on Natural Language Generation (INLG19) で発表したものである (Nishimura, Hashimoto, and Mori 2019)．

参考文献

Balntas, V., Riba, E., Ponsa, D., and Mikolajczyk, K. (2016). "Learning local feature descriptors with triplets and shallow convolutional neural networks." In *Proceedings of the British Machine Vision Conference*, pp. 1–11.

Biten, A. F., Gomez, L., Rusinol, M., and Karatzas, D. (2019). "Good news, everyone! context driven entity-aware captioning for news images." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 12466–12475.

Bosselut, A., Celikyilmaz, A., He, X., Gao, J., Huang, P.-S., and Choi, Y. (2018). "Discourse-Aware Neural Rewards for Coherent Text Generation." In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 173–184.

Chandu, K., Nyberg, E., and Black, A. W. (2019). "Storyboarding of recipes: grounded contextual generation." In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 6040–6046.

Chen, J. and Ngo, C.-W. (2016). "Deep-based Ingredient Recognition for Cooking Recipe Retrieval." In *Proceedings of the ACM International Conference on Multimedia*, pp. 32–41.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Harashima, J., Someya, Y., and Kikuta, Y. (2017). "Cookpad image dataset: An image collection as infrastructure for food research." In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1229–1232.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Ishan Misra, A. A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. (2016). "Visual Storytelling." In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239.

Kiddon, C., Zettlemoyer, L., and Choi, Y. (2016). "Globally Coherent Text Generation with Neural Checklist Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 329–339.

Kingma, D. P. and Ba, J. (2015). "Adam: A method for stochastic optimization." In *Proceedings*

of the International Conference for Learning Representations.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). "Microsoft coco: Common objects in context." In *Proceedings of the European Conference on Computer Vision*, pp. 740–755.

Liu, Y., Fu, J., Mei, T., and Chen, C. W. (2017). "Let Your Photos Talk: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks." In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1445–1452.

Luong, T., Pham, H., and Manning, C. D. (2015). "Effective Approaches to Attention-based Neural Machine Translation." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421.

Malmaud, J., Wagner, E. J., Chang, N., and Murphy, K. (2014). "Cooking with semantics." In *Proceedings of the ACL Workshop on Semantic Parsing*, pp. 33–38.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality." In *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Mori, S., Maeta, H., Sasada, T., Yoshino, K., Hashimoto, A., Funatomi, T., and Yamakata, Y. (2014a). "FlowGraph2Text: Automatic sentence skeleton compilation for procedural text generation." In *Proceedings of the International Conference on Natural Language Generation*, pp. 118–122.

Mori, S., Maeta, H., Yamakata, Y., and Sasada, T. (2014b). "Flow Graph Corpus from Recipe Texts." In *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 2370–2377.

Neubig, G., Nakata, Y., and Mori, S. (2011). "Pointwise prediction for robust, adaptable Japanese morphological analysis." In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 529–533.

Nishimura, T., Hashimoto, A., and Mori, S. (2019). "Procedural text generation from a photo sequence." In *Proceedings of the International Conference on Natural Language Generation*, pp. 409–414.

Salvador, A., Drozdzal, M., Giro-i-Nieto, X., and Romero, A. (2019). "Inverse cooking: Recipe generation from food images." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 10453–10462.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., and Torralba, A. (2017). "Learning Cross-modal Embeddings for Cooking Recipes and Food Images." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 3020–3028.

Ushiku, A., Hashimoto, H., Hashimoto, A., and Mori, S. (2017). "Procedural text generation from an execution video." In *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 326–335.

Wang, L., Li, Y., Huang, J., and Lazebnik, S. (2016). "Learning Deep Structure-Preserving Image-Text Embeddings." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 5005–5013.

Yagcioglu, S., Erdem, A., Erdem, E., and Ikizler-Cinbis, N. (2018). "RecipeQA: a challenge dataset for multimodal comprehension of cooking recipes." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1358–1368.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). "Image captioning with semantic attention." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.". **2**.

Zhu, B., Ngo, C.-W., Chen, J., and Hao, Y. (2019). "R2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network." In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 11477–11486.

笹田鉄郎，森信介，山肩洋子，前田浩邦，河原達也 (2015). レシピ用語の定義とその自動認識のためのタグ付与コーパスの構築. 自然言語処理, **22** (2).

略歴

西村 太一：平 31 九州大学芸術工学部卒業．京都大学大学院情報学研究科修士課程在籍中．マルチメディア，自然言語処理の研究を手掛ける．令元年 The 11th Workshop on Multimedia for Cooking and Eating Activities にて，Best Paper Award を受賞．

橋本 敦史：平 17 京大．工・情報卒．平 18 年経産省 Vulcanus in Europe プログラム国費奨学生．平 25 京大大学院情報学研究科にて博士 (情報学) 取得．現在オムロンサイニックエックス株式会社研究員．主に，料理や組立作業を対象として，未来予測に基づく人と機械のインタラクションに関する研究などに従事．IEEE, IEICE, IPSJ 各会員．

森 信介：平 5 京大．工・電気卒．平 7 京大．工・修士修了．平 10 京大．工・博士後期課程修了．同年日本アイ・ビー・エム株式会社入社．平 19 より京都大学学術情報メディアセンター准教授．平 28 同教授．現在に至る．計算言語学ならびに自然言語処理の研究に従事．工学博士．平 9 情報処理学会山下記念

研究賞受賞．平 22, 平 25 情報処理学会論文賞受賞．平 22 第 58 回電気科学技術奨励賞．言語処理学会，情報処理学会，日本データベース学会各会員．