社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

作業写真列からの手順書の自動生成

西村　太一† 　橋本　敦史†† 　森　　信介†††

† 京都大学大学院 情報学研究科
〒606-8501 京都府京都市左京区吉田本町
†† オムロンサイニックエックス株式会社
〒113-0033 東京都文京区本郷 5 丁目 24 － 5 角川本郷ビル 3F
††† 京都大学学術情報メディアセンター
〒606-8501 京都府京都市左京区吉田本町
E-mail: †nishimura.taichi.43x@st.kyoto-u.ac.jp, ††atsushi.hashimoto@sinicx.com, †††forest@i.kyoto-u.ac.jp

あらまし　本研究は，写真列を入力として手順書を生成し，写真付き手順書の作成を容易にすることを目的とする．この目的を達成するために，モデルは写真を説明する上で欠かせない物体や動作の表現 (重要語) を含んだ手順書を生成することが求められる．従来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，本研究では検索問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む手法を提案する．これにより，モデルは入力写真に適した重要語を含む手順を検索し，参照しながら単語を出力することで，写真に適した重要語を含んだ手順書を生成することができる．実験では，料理タスクを対象に手順書生成を行なった．その結果，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確認できた．
キーワード　手順書，写真列，共有潜在空間，文生成

Procedural Text Generation from a Photo Sequence

Taichi NISHIMURA†, Atsushi HASHIMOTO††, and Shinsuke MORI†††

† Graduate School of Informatics, Kyoto University
Yoshidahonmachi, Sakyoku, Kyoto, Japan, 606-8501
†† OMRON SINIC X Corporation
24-5, Hongo, Bunkyo-ku, Tokyo, Japan, 113–0033
††† Academic Center for Computing and Media Studies, Kyoto University
Yoshidahonmachi, Kyoto-ku, Kyoto, Japan, 606–8501
E-mail: †nishimura.taichi.43x@st.kyoto-u.ac.jp, ††atsushi.hashimoto@sinicx.com, †††forest@i.kyoto-u.ac.jp

Abstract　In this paper, we tackle a problem to generate a procedural text from a photo sequence, which aims to help users create a multimedia procedural text only by taking photographs. For this goal, the output texts should include important words that make sense as an instructions. However, traditional methods do not consider these words. To select the important words to describe a photo, the proposed method incorporates a retrieval method into a generation model. From various experimental results, we confirmed that the method outperforms standard baselines.
Key words　Procedural text, Photo sequence, Cross modal embedding, Sentence generation

1. はじめに

言語による各指示に対して，指示の内容を示す視覚的情報が付与された「写真付き手順書」があると，読者は作業内容を理解しやすくなる．しかし，写真付き手順書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手順書を自動生成することで，写真付き手順書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

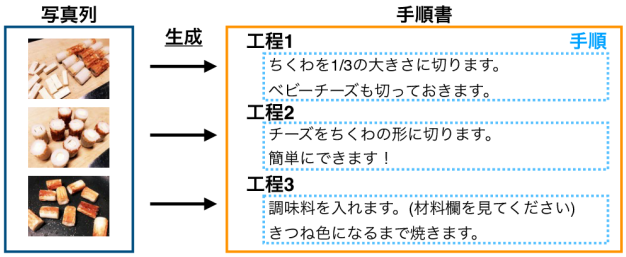

図 1　写真列からの手順書の自動生成．入力が写真列であり (左)，出力が複文からなる手順である (右)．手順は写真列の各写真ごとに生成する．

ムは各写真ごとに複数の文からなる手順を生成する問題 (図 1) として定式化し，この問題を解決する手法を提案する．

この問題設定は入出力が共通しているという点で，visual storytelling [1] に類似している．しかし，visual storytelling と比較して，出力の手順書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，対象とするタスクにおける重要な物体や動作を含んだ表現が正しく生成されなければならない．これらを表現する対象のタスクに固有に現れる単語を，本論文では重要語と定義する．本論文で対象とする料理タスクでは，食材，道具，調理者の動作を表す単語がこれに該当する．

このような重要語を含んだ手順書を生成するアイデアとして，各写真に適した重要語を含む手順を検索し，それらの表現を参照しながら手順を生成することが考えられる．このアイデアを実現するために，本論文では検索問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むことで，写真列に適した重要語を含む手順書を生成する手法を提案する．提案手法では，2 段階に分けて手順書を生成する．最初に，写真列の各写真ごとに，あらかじめ大規模な写真と手順の組で学習した写真と手順の共有潜在空間を用いて写真に適した手順を検索する．次に，得られた手順を参照しながら文生成を行い，写真に適した重要語を含む手順を生成する．

本手法を実装し，料理タスクを対象に評価実験を行なった．実験では，提案手法を visual storytelling のベースラインと比較し，文生成の定量的評価尺度である BLEU，ROUGE-L，CIDEr-D において性能向上を示した．これらの自動評価尺度に加えて，我々は重要語が正しく生成文に現れている割合を調査した結果，提案手法が正しく重要語を含んだ手順を生成していることが確認できた．

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，本論文の問題設定は visual storytelling [1] と類似している．visual storytelling の課題に対して，Liu ら [8] は画像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含む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，重要語を生成しながら手順書を生成することができる．

手順書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．中でも，料理という題材は，Web 上でデータを集めやすく，手順書の記述の表現が多様である．そのため，料理タスクで利用可能な手法を構築できれば，他の多くの手順書を生成するタスク (例: 家具の組み立て方など) に適用できるため，活発に研究されている．Salvador ら [4] 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，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 [6] はレシピのタイトルと食材を入力として与え，生成文で材料を利用したかどうかを注意機構 [11] 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，本研究では食材に加えて調理者の動作や道具も重要語として考慮する点が異なる．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手順書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，十分な精度で手順書を生成するまでには至っていない．一方で，Mori ら [7] 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 [17] を入力として手順書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，実用的な精度で手順書を生成することに成功している．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．写真の撮影者は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，入力に手順途中の情報が十分に含まれている．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，より高い精度で手順書を生成することが期待できる．

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含む手順書を生成する手法について説明する．本論文では，検索問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，明示的に重要語を参照しながら手順書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)　あらかじめ，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学習させておく．その後，入力された写真列をもとに，写真列の各写真ごとに以下の (ii) から (iv) の手続きを繰り返して手順を生成し，生成した全ての手順をあわせて手順書として出力する．

(ii)　写真に適した重要語を含む手順ベクトルを得るために，共有潜在空間上において入力写真に近傍の 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双方向 LSTM (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．

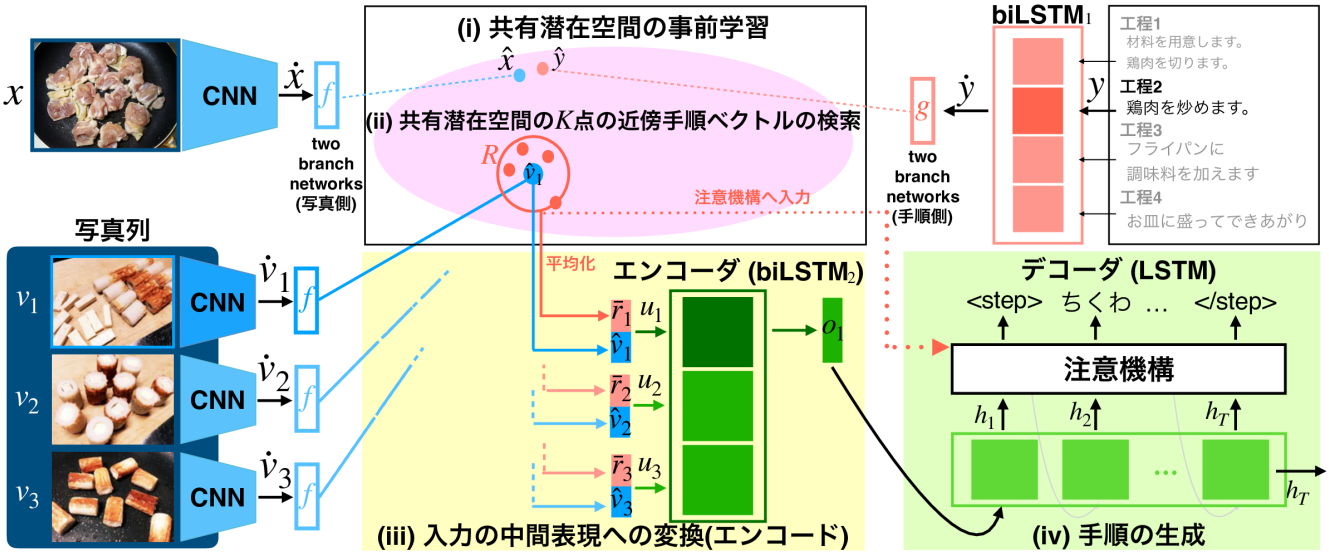

図 2　提案手法の概要.

3.1　共有潜在空間

最初に，(i) の学習を行うために，Wang ら [9] 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，類似する写真と手順のペアは共有潜在空間上で近く，類似しない場合は遠くなるように位置している．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，手順書中では，非常に多く物体名が省略される [10]．そのため，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 $\dot{\boldsymbol{x}}$，手順中の各単語を word2vec [15] 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 $\boldsymbol{y}$ 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 $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ は以下のように表される．

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

$$\dot{\boldsymbol{y}} = \mathrm{biLSTM}_1(\boldsymbol{y}),\ \hat{\boldsymbol{y}} = g(\dot{\boldsymbol{y}}) \tag{2}$$

ここで，$\mathrm{CNN}(\cdot)$ は図中の CNN に対応し，$\mathrm{biLSTM}_1(\cdot)$ は追加した biLSTM である図中の biLSTM_1 に対応する．また，$\dot{\boldsymbol{y}}$ は biLSTM の出力ベクトルを表す．

3.2　手順書生成

入力の写真列を $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N)$ 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．こうして得られる写真の埋め込みベクトル $\hat{\boldsymbol{v}}_n$ は以下のように表される．

$$\dot{\boldsymbol{v}}_n = \mathrm{CNN}(\boldsymbol{v}_n),\ \hat{\boldsymbol{v}}_n = f(\dot{\boldsymbol{v}}_n) \tag{3}$$

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:
画像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．

$$\bar{\boldsymbol{r}}_n = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{r}_k \tag{4}$$

最後に，得られた手順ベクトルの平均ベクトルと，画像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．

$$\boldsymbol{o}_n = \mathrm{biLSTM}_2(\boldsymbol{u}_n) \tag{5}$$

ここで，$\mathrm{biLSTM}_2(\cdot)$ は図中の 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 (⟨step⟩) から終端記号 (⟨/step⟩) が生成されるまで，単語を 1 つ 1 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 [11] をモデルに組

		訓練	検証	評価
D_{emb}	レシピ数	162,463	18,059	20,104
	工程数	5.65	5.57	5.66
	単語数	24.51	24.51	24.40
	語彙サイズ		24,152	
D_{gen}	レシピ数	21,039	2,281	2,598
	工程数	8.09	8.10	8.10
	単語数	19.35	19.51	19.32
	語彙サイズ		11,091	

表 1　データセットの統計結果.

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，重要語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．t 番目の単語の出力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 を用いて，t 番目の単語を出力する時の，k 個目の手順ベクトルへの注意機構の重み a_k^t，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．

$$a_k^t = \frac{\exp\left(\boldsymbol{r}_k \boldsymbol{W}_a \boldsymbol{h}_t\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{r}_j \boldsymbol{W}_a \boldsymbol{h}_t\right)} \tag{6}$$

$$\boldsymbol{c}_t = \sum_{k=1}^{K} a_k^t \boldsymbol{r}_k \tag{7}$$

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_c(\boldsymbol{c}_t, \boldsymbol{h}_t)) \tag{8}$$

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．

$$p(y_t|y_{<t}, \boldsymbol{o}_n) = \mathrm{softmax}(\boldsymbol{W}_o \tilde{\boldsymbol{h}}_t + \boldsymbol{b}_o) \tag{9}$$

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．学習を行うときは，学習データの全てにおいて，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．

$$L(\boldsymbol{\theta}) = -\sum_{\mathcal{D}} \sum_{t=1}^{T} \log p(y_t|y_{<t}, \boldsymbol{o}_n; \boldsymbol{\theta}) \tag{10}$$

ここで，$\mathcal{D}$ は学習データセット全体を，$\boldsymbol{\theta}$ は全ての学習可能な重みを表し，T は出力する手順の単語数を表す．

4. 実験と評価

提案手法を実装し，料理タスクを対象に評価実験を行った．料理タスクは他の手順書のタスクと比較し，手順書の記述の表現が多様であり，料理タスクで手法を構築できれば他のタスクにも応用できると考えられる．そのため，本研究で取り組むタスクとして相応しいと考え，選択した．

	image2step	step2image
biLSTM$_1$(なし)	23	24
biLSTM$_1$(あり)	**6**	**6**

表 2　two branch networks [9] へ biLSTM を追加したときの medR の結果.

4.1 詳細設定

共有潜在空間への画像側のエンコーダとして，ImageNet [13] で学習済みの ResNet-50 [12] 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，画像側の出力の次元数は 2,048 である．共有潜在空間でのテキスト側のエンコーダの biLSTM 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 と同様のプロセスで学習を行なった [9]．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 [14] 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．

4.2 データセット

実験で利用するデータセットは, Cookpad Image Dataset [16] 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(注1)．写真列に含まれる写真数によるモデルへの影響を減らすために，7 手順以上，10 手順以下の手順書のみを抽出し，これを手順書生成のデータセット D_{gen} とした．残りのデータセットを D_{emb} として共有潜在空間の学習のためのデータセットとした．また，出現頻度数が 3 回以下の単語は未知語とした．表 1 にデータセットの統計情報を示す．

4.3 共有潜在空間への biLSTM の影響

最初に，two branch networks へ biLSTM 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，ランダムに D_{emb} のテストセットから 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) 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 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．これは，biLSTM 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であると考えられる．

(注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．

		BLEU1	BLEU4	ROUGE-L	CIDEr-D
ベースライン	写真列	27.3	4.2	18.3	13.2
	写真列 + タイトル	28.6	5.4	17.6	13.1
	写真列 + タイトル + 材料	28.8	6.1	19.4	14.6
提案手法 biLSTM$_1$(なし)	写真の埋め込みベクトル + Top1 の手順ベクトル	26.7	4.1	17.7	13.8
	写真の埋め込みベクトル + Top$K$ の手順ベクトル	31.4	6.8	21.5	11.7
提案手法 biLSTM$_1$(あり)	写真の埋め込みベクトル	31.0	6.5	21.6	14.9
	写真の埋め込みベクトル + Top1 の手順ベクトル	32.9	6.7	**21.8**	**16.4**
	写真の埋め込みベクトル + Top$K$ の手順ベクトル	**33.4**	**7.2**	20.7	14.9

表 3　自動評価尺度による生成した手順書の評価結果．実験では，$K = 10$ として設定した．

図 3　手順書の生成例．太字で書かれた箇所は正しく手順が生成された箇所であり，下線部付きで書かれた箇所は正解と比較して生成文として不正な箇所である．また，太字の箇所のうち，二重線付きの箇所は正しく手順の重要語が生成できた箇所である．

4.4 結　　果

提案手法を評価するために，BLEU, ROUGE-L, CIDEr-D といった文生成の自動評価尺度を評価するとともに，参照文中の重要語が，生成した手順書中に正しく現れたかどうかも評価した．また，実際に写真列から生成した手順書の一例を示し，提案手法の有用性を確認した．

4.4.1 定量的評価

提案手法を評価するために，テストセットの全ての手順書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．検索問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 [1] で挙げられている，写真列をエンコーダの LSTM 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，手順書に付与した材料やタイトルの各単語を word2vec [15] 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)．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．

4.4.2 重要語の生成率

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，正解の手順書の重要語が，生成した手順

		F	T	Ac	合計
ベースライン	再現率	7.9	22.6	19.2	14.8
	適合率	12.3	15.8	17.0	15.4
	F 値	9.6	18.6	18.0	15.1
提案手法 (Top1) biLSTM$_1$(あり)	再現率	18.5	24.7	31.6	25.2
	適合率	23.8	21.0	21.1	21.9
	F 値	20.8	22.7	25.3	23.4
提案手法 (Top$K$) biLSTM$_1$(あり)	再現率	40.5	29.8	35.9	37.2
	適合率	43.6	26.8	32.4	36.1
	F 値	**42.0**	**28.2**	**34.0**	**36.6**

表 4　重要語を正しく生成できた割合．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．

書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．料理ドメインにおいては，レシピフローグラフコーパス [17] 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，適切に生成した手順書にこれらが現れているかどうかを測定することで，重要語の生成率を評価した．なお，重要語の生成率の評価尺度として，以下で表される適合率，再現率，F 値を用いた．

$$\text{再現率} = \frac{\text{正解の重要語数}}{\text{参照文中の手順書に現れる重要語数}} \tag{11}$$

$$\text{適合率} = \frac{\text{正解の重要語数}}{\text{生成した手順書に現れる重要語数}} \tag{12}$$

$$F\text{ 値} = \left(\frac{\text{再現率}^{-1} + \text{適合率}^{-1}}{2}\right)^{-1} \tag{13}$$

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，手順書を 50 個ランダムでテストセットから抽出し，手動で同義語と表記揺れのみを修正し，重要語の生成率を計算した．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(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(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手順書を生成していると言える．

4.4.3　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手順書，そして正解の手順書を載せる．この図より，ベースラインは写真に適した手順書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含んで生成しているということが分かる．

5. ま と め

本論文では，写真付き手順書の作成を支援するために，写真列から手順書を生成する問題として定式化し，写真列に適した重要語を含む手順書を生成する手法を提案した．そして，本手法を実装し，料理ドメインを対象に評価実験を行なった．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，重要語が正しく生成できているかと言う点においても visual storytelling の標準的なベースラインを上回ることを実験的に示した．また，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．以上の結果より，本論文の提案手法を用いて写真列に適した手順書を得ることができていることが実験的に確認できた．よって，本研究で構築したシステムは，作者の写真付き手順書生成の支援に応用することができると考えられる．

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ました．よってここに感謝の意を表します．

文　献

[1] Huang, Ting-Hao Kenneth et al. Visual Storytelling. In *Proc. of NAACL*, pp.1233-1239, 2016.
[2] Amaia Salvador et al. Learning Cross-modal Embeddings for Cooking Recipes and Food Images. In *Proc. of CVPR*, pp.3020-3028, 2017.
[3] Bin Zhu et al. R2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network. In *Proc. of CVPR*, pp.11477-11486, 2019.
[4] Amaia Salvador et al. Inverse cooking: Recipe generation from food images. In *Proc. of CVPR*, pp.10453-10462, 2019.
[5] Antoine Bosselut et al. Discourse-Aware Neural Rewards for Coherent Text Generation. In *Proc. of NAACL*, pp.173-184, 2018.
[6] Chloé Kiddon et al. Globally Coherent Text Generation with Neural Checklist Models. In *Proc. of EMNLP*, pp.329-339, 2016.
[7] Shinsuke Mori et al. FlowGraph2Text: Automatic sentence skeleton compilation for procedural text generation. In *Proc. of INLG*, pp.118-122, 2014.
[8] Yu Liu et al. Let Your Photos Talk: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks. In *Proc. of AAAI*, pp.1445-1452, 2017.
[9] Liwei Wang et al. Learning two-branch neural networks for image-Text matching tasks. In *Proc. of CVPR*, pp.5005-5013, 2016.
[10] Jonathan Malmaud et al. Cooking with semantics. In *Proc. of ACL Workshop on Semantic Parsing*, pp.33-38, 2014.
[11] Thang Luong et al. Effective Approaches to Attention-based Neural Machine Translation. In *Proc. of EMNLP*, pp.1412-1421, 2015.
[12] Kaiming He et al. Deep residual learning for image recognition. In *Proc. of CVPR*, pp.770-778, 2016.
[13] Jia Deng et al. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*, pp.248-255, 2009.
[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.
[15] Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.
[16] Jun Harashima et al. Cookpad image dataset: An image collection as infrastructure for food research. In *Proc. of SIGIR*, pp. 229-1232, 2017.
[17] Shinsuke Mori et al. Flow Graph Corpus from Recipe Texts. In *Proc. of LREC*, pp. 2370-2377, 2014.