画像と単語の不一致を考慮した疑似教師ありキャプション生成

本多 右京 [1,3]　牛久 祥孝 [2]　橋本 敦史 [2]　渡辺 太郎 [1]　松本 裕治 [3]
[1] 奈良先端科学技術大学院大学大学　[2] オムロンサイニックエックス株式会社
[3] 理化学研究所
{honda.ukyo.hn6, taro}@is.naist.jp
{yoshitaka.ushiku,atsushi.hashimoto}@sinicx.com
yuji.matsumoto@riken.jp

1 はじめに

画像キャプション生成は，画像から自然言語で説明文を生成するタスクである．人手でアノテーションされた膨大な画像とキャプションのペア [1, 2] を学習データとして，目覚ましい進歩が報告されてきた [3, 4, 5]．しかし，大規模な学習データでもカバーする事物は限定的であり[1)]，記述できる事物を増やすためには更にアノテーションコストがかかる．

疑似教師ありキャプション生成[2)]は，画像とキャプションのペアがない事物に対して，アノテーションなしでキャプション生成を行う試みである [8, 9]．学習に画像とキャプションのペアを用いないことでこの状況を作り，物体検出モデルの出力した物体名と，アノテーションが不要で大量に利用可能な，ペアでない画像と文だけを入力として学習を行う．

先行研究 [8, 9] は，画像から検出された物体名を含んだ文を**擬似キャプション**とし，画像の特徴量とこの一文全体の特徴量とを近づける学習を行う．しかし，疑似キャプションには画像に対応しない記述が多く含まれるという問題がある．例えば，cat と girl が検出された画像に対する疑似キャプションには “a girl is holding a cat” や “a cat is sleeping with a girl”, “a girl is running with a cat” などが考えられるが，最初の文が正しく画像を説明した記述だった場合，それ以外の文の “sleeping with” や “running with” といった記述は画像に対して誤った教師信号を与えてしまうことになる（図 1 左）．

この部分的な記述の不一致に対処するため，本研究では，簡単なゲート機構と記述の一致・不一致に対する疑似教師信号を用いて，画像と齟齬のある単語については画像特徴量から生成しないよう抑制する手法を提案する．実験の結果，提案手法が先行研究の性能を上回り，部分的な不一致を考慮することの重要性が示された．また，提案手法は，モデルの初期化手法として先行研究の手法と組み合わせることで，さらに性能を向上させることが確認された．

2 関連研究

学習データにない事物を記述する試みとしては，学習データに現れない物体を物体検出の出力を利用して記述する novel object captioning [10, 6, 11] や，物体，属性，関係に関して学習データにない組み合わせの事例でテストを行う試み [12, 13] などがある．

これらのタスクは，ペアになった画像とキャプションがテストデータの大部分の要素をカバーする状況でペアを学習に用いる．疑似教師ありキャプション生成はこの状況が成り立たないテストを想定し，ペアになっていない画像と文だけを用いて学習を行う [8]．ただし，novel object captioning と同様，事前学習した物体検出器で予測された物体名は利用する．この準備段階的なタスクである unpaired image captioning [8, 9, 14, 15] では，扱う画像と文は元々ペアになっていたものだけを使うのに対して，疑似教師ありキャプション生成では画像と文はそれぞれ別のリソースから取得したものを想定する．

先行研究 [8, 9] の問題点は，いずれも擬似キャプションの一文全体の特徴量を画像特徴量と対応付けるよう学習するため，1 節で指摘したように，疑似キャプション中の部分的な不一致も学習に使ってし

1) 代表的なデータセットである MS COCO[2] では，物体検出データに定義されている 500 の物体カテゴリのうち約 400 がキャプションに一度も出現しない [6]．また，学習に用いられる語彙数が全体で 8791 と少なく [7]，物体以外についてもカバーしている事物が限定的である．

2) 先行研究 [8, 9] では，画像とキャプションのペアが要求されないという意味で，**教師なしキャプション生成**と呼ばれているタスクである．しかし，この名称では物体検出モデルを要求する点に誤解が生じるおそれがあるため，本稿ではこのタスクを疑似教師ありキャプション生成と呼称する．

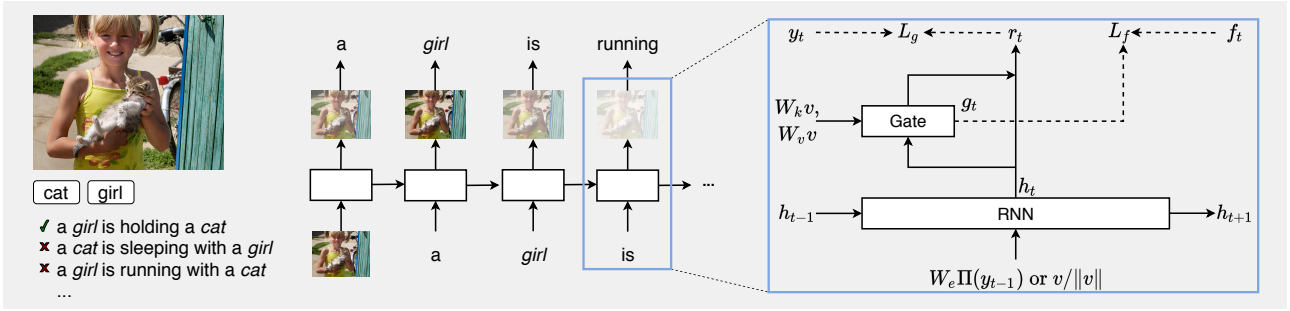

図 1 提案手法概要．モデルへの入力は，画像，検出された物体名，疑似キャプション（図左）．モデルは疑似キャプション中の各単語を，画像との対応度を考慮しながら生成するよう学習する（図右）．基本モデルの出力 h，ゲートの値 g，ゲートへの疑似教師信号 f の詳細はそれぞれ，3.1，3.2，3.3 節を参照．破線は学習時のみ適用される処理を表す．

まうことである．[8] はまず，物体名を含む任意の文に対して，その含まれる物体名のエンコードから元の文を復元する疑似キャプション生成器を学習する．この疑似キャプション生成器を画像から検出された物体名に対して適用し，その出力をその画像に対する疑似キャプションとする．[9] では，画像から検出された物体名を含む文をそのまま疑似キャプションとして用いる．疑似キャプションを得た後は，画像と擬似キャプションの特徴量に対して，双方向の特徴量再構成 [8] や距離学習 [9] を適用して，画像と擬似キャプションとの対応付けを学習する．

3 手法

本研究では，画像と齟齬のある単語を画像に対する教師信号として用いないよう抑制する手法を提案する．手法の概要を図 1 に示す．

3.1 基本モデル

教師ありキャプション生成では画像 $\boldsymbol{I}$ に対して正解のキャプション $\boldsymbol{y}$ が与えられるが，疑似教師ありキャプション生成では，画像から検出された物体名を含む任意の文 $\hat{\boldsymbol{y}}$ で $\boldsymbol{y}$ を代替する．提案手法の基本となるモデルでは，画像 $\boldsymbol{I}$ とその疑似キャプション $\hat{\boldsymbol{y}} = [\hat{y}_1, ..., \hat{y}_T]$（$\hat{y}_T$ は<eos>トークン）の各ペアに対して，条件付き確率 $p(\hat{\boldsymbol{y}} \mid \boldsymbol{I}) = \prod_{t=1}^{T} p(\hat{y}_t \mid \hat{y}_{<t}, \boldsymbol{I})$ を以下のようにモデル化する：

$$p(\hat{y}_t \mid \hat{y}_{<t}, \boldsymbol{I}) = \frac{\exp(\boldsymbol{h}_t^\top \boldsymbol{W}_o \Pi(\hat{y}_{t-1}))}{\sum_{\hat{y}} \exp(\boldsymbol{h}_t^\top \boldsymbol{W}_o \Pi(\hat{y}))}, \tag{1}$$

$$\boldsymbol{h}_t = \begin{cases} \mathrm{Dec}\left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}, \boldsymbol{h}_0\right) & \text{if } t = 1 \\ \mathrm{Dec}(\boldsymbol{e}_t, \boldsymbol{h}_{t-1}) & \text{otherwise.} \end{cases} \tag{2}$$

入力は，$\boldsymbol{v} = \boldsymbol{W}_a \mathrm{Enc}(\boldsymbol{I})$，$\boldsymbol{e}_t = \boldsymbol{W}_e \Pi(\hat{y}_{t-1})$，ゼロベクトル $h_0 \in \mathbb{R}^d$．ここで，$\Pi(x)$ は単語 x に対応する one-hot ベクトルを出力する関数，$\mathrm{Dec}(\cdot)$ は RNN，$\mathrm{Enc}(\cdot)$ は画像分類を事前学習した CNN，$\boldsymbol{W}_a \in \mathbb{R}^{d \times d'}$ は画像特徴量に対する線形変換行列，$\boldsymbol{W}_e, \boldsymbol{W}_o \in \mathbb{R}^{d \times N}$ は語彙数 N の単語埋め込み行列である．

3.2 ゲート機構の導入

疑似キャプションは，1 節で述べたように，画像に対応しない記述を含む．この画像と擬似キャプションとの部分的な不一致を考慮するため，疑似キャプション中の各単語の生成に画像の特徴量をどの程度使うかを制御するゲート機構を導入する．ゲート機構は式 (1) を以下のように変更する：

$$p(\hat{y}_t \mid \hat{y}_{<t}, \boldsymbol{I}) = \frac{\exp(\boldsymbol{r}_t^\top \boldsymbol{W}_o \Pi(\hat{y}_{t-1}))}{\sum_{\hat{y}} \exp(\boldsymbol{r}_t^\top \boldsymbol{W}_o \Pi(\hat{y}))}, \tag{3}$$

$$\boldsymbol{r}_t = g_t \frac{\boldsymbol{W}_v \boldsymbol{v}}{\|\boldsymbol{W}_v \boldsymbol{v}\|_2} + (1 - g_t)\boldsymbol{h}_t, \tag{4}$$

$$g_t = \mathrm{sigmoid}(\tanh(\boldsymbol{W}_k \boldsymbol{v})^\top \boldsymbol{h}_t). \tag{5}$$

ここで，$g \in [0, 1]$ はゲートの値，$\boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ は画像特徴量に対する線形変換行列である．ゲートを通すことで，各タイムステップでの出力で画像特徴量を g の重み付きで用いることができる．これにより，疑似キャプション中で画像に対応する部分については g を大きく，そうでない部分については g を小さくして，画像と疑似キャプションとの部分的な一致・不一致を考慮することを可能にする．

画像と擬似キャプションの各ペアに対して，ゲート機構付きモデルは以下の交差エントロピー誤差 L_g を最小化するよう学習する：

$$L_g = -\frac{1}{T} \sum_{t=1}^{T} \log p(\hat{y}_t \mid \hat{y}_{<t}, \boldsymbol{I}). \tag{6}$$

3.3 ゲートへの疑似教師信号

3.2 節で導入したゲート機構は，疑似キャプション中でどの単語が画像に対応しているのかについて教師信号を与えられていない．タスクの設定上正し

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Prev. [8]	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
Ours	**49.5 ± 0.7**	**27.3 ± 1.2**	**13.1 ± 0.8**	**6.3 ± 0.5**	**14.0 ± 0.1**	**34.5 ± 0.3**	**31.9 ± 1.0**	**8.6 ± 0.2**
Prev. [9]				6.5	12.9	35.1	22.7	
Ours	50.4 ± 1.5	29.5 ± 0.8	14.4 ± 0.5	**7.6 ± 0.4**	**13.5 ± 0.3**	**37.3 ± 0.2**	**31.8 ± 0.7**	8.4 ± 0.1

表 1　先行研究との比較．二重線の上下は，それぞれ先行研究 Prev. [8, 9] の実験設定に対応する．提案手法のスコアは 5 回の試行の平均と標準偏差．Prev. [9] の BLEU-1 から 3 と SPICE のスコアについては，論文での報告がないため空欄．

い教師信号を与えることはできないが，以下のように擬似的に教師信号を与えることはできる．

疑似キャプションは，画像から検出された物体名だけを頼りに取得される．そのため，物体名以外の単語については画像と対応している可能性が低い．提案手法では，この対応関係を反映した擬似教師信号 f を用いる．f は，$\hat{y}_t$ が画像から検出された物体名である場合は $f_t = 1$，それ以外の場合は $f_t = 0$ となる．この f を g に対する教師信号として，以下の交差エントロピー誤差 L_f を最小化する：

$$L_f = -\frac{1}{T}\sum_{t=1}^{T}\{\alpha f_t \log g_t + (1 - f_t)\log(1 - g_t)\}. \quad (7)$$

ここで，α は $f_t = 1$ での誤差を重み付けるハイパーパラメータである．疑似キャプション中で検出物体名に対応する単語はそれ以外の単語に比べて数が少ないため，α にはこの比率を考慮した値を設定する．

ゲートに上記の疑似教師信号を与えると，式 (4) から，基本的に，物体名については画像特徴量を使って生成し，それ以外の単語については前タイムステップまでの文脈特徴量を使って生成することになる．これにより，物体名以外については画像特徴量に対応した生成を学習できなくなる反面，画像に一致しない可能性の高い記述を画像から生成するよう強制してしまうことが避けられる．

提案手法の最終的な損失関数 L は以下とする：

$$L = L_g + L_f. \quad (8)$$

3.4　物体名デコードへの制約

式 (4) では，$g_t = 1$ のとき，t によらず同じ特徴量 $\boldsymbol{W}_v\boldsymbol{v}/\|\boldsymbol{W}_v\boldsymbol{v}\|_2$ から $\hat{y}_t$ が予測される．このため，このまま用いるとデコード時に同じ物体名を繰り返し出力する問題が起きる．そこでデコード時には，物体検出器に定義されている物体カテゴリのうち，一度生成した物体名についてはそれ以降のタイムステップでの生成確率を 0 にする制約を加える．

		Precision	Recall	F1
物体名	Prev. [8]	**37.5**	36.3	**34.3**
	Ours	26.8	**38.8**	29.7
その他	Prev. [8]	24.3	17.6	19.7
	Ours	**34.0**	**22.7**	**26.0**

表 2　出力キャプションとその参照キャプションから抽出した，物体名 / その他の単語の BoW の平均一致度．

4　実験

4.1　実験設定概要

先行研究に従って，学習に使う画像データには MS COCO [2] の学習データ [7]，テキストデータには Web から取得された画像説明文 [8, 16] を使用した．ただし，これらのテキストデータには MS COCO のキャプションは含まれていない．評価は，MS COCO の評価用データに対して BLEU [17]，ROUGE [18]，METEOR [19]，CIDEr [20]，SPICE [21] の自動評価指標を適用して行った．表の数値はすべて，MS COCO のテストデータで算出した．

実験設定の詳細部分では，先行研究 [8, 9] はそれぞれやや異なる設定をとっている．詳細については付録 A を参照されたい．

4.2　先行研究との比較

表 1 に先行研究のスコアとの比較を示す．先行研究のいずれの設定においても，すべての評価指標で提案手法が先行研究を上回るスコアを示しており，提案手法の有効性が確認された．

4.3　物体名以外の生成における比較

疑似キャプション中で画像との不一致がある可能性が高いのは物体名以外の単語である（3.3 節）．提案手法はこの不一致を学習に使わせないことに重点

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Prev. [8]	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
Ours	49.5 ± 0.7	**27.3 ± 1.2**	**13.1 ± 0.8**	6.3 ± 0.5	**14.0 ± 0.1**	34.5 ± 0.3	31.9 ± 1.0	8.6 ± 0.2
Ours + Prev. [8]	**50.9 ± 0.1**	**28.0 ± 0.1**	**14.0 ± 0.1**	**7.1 ± 0.0**	**14.1 ± 0.0**	**35.2 ± 0.1**	**35.7 ± 0.1**	**9.2 ± 0.0**

表 3　提案手法と先行研究手法 Prev. [8] の併用の結果．Ours が含まれる手法のスコアは 5 回の試行の平均と標準偏差．

を置くので，特に物体名以外の単語の生成において先行研究に対する改善があることが期待される．この改善を確かめるため，出力キャプションとその参照キャプションから，物体検出器に定義されている物体カテゴリの単語とそれ以外の単語のそれぞれを抽出した Bag of Words (BoW) を作成し，出力-参照ペアで BoW の平均一致度を算出した．比較は，実装が公開されている先行研究 [8] に対して行った．

表 2 から，提案手法が物体名以外の単語の生成で先行研究 [8] を上回っていることがわかる．特に Precision での差が大きいことから，学習において特に画像との不一致が多い部分については，これを画像に対して教師信号として与えると，誤って画像に対応付けられた余計な記述が生成されてしまうのだと考えられる．これに対して提案手法は，物体名以外は基本的に前文脈に対して頻度の高い単語を生成する学習を行う．これによる頻出系列の方が，画像に一致した出力になりやすいことがわかった．

4.4　モデルの初期化手法としての活用

提案手法が画像に対してそれと齟齬のある単語を対応付けない学習に重点を置いているのに対して，先行研究は多少の齟齬はあっても画像に一文全体を対応付ける学習に重点を置いている．両者が重視している互いに異なる要素を補うため，これらを組み合わせた手法の実験を行った．先行研究 [8] の実験設定で，学習データの画像に対して学習済みの提案モデルでキャプションを出力し，画像とキャプションのペアを作成する．このペアを使って画像から上記のキャプションを生成する学習を行い，先行研究のキャプション生成モデルのパラメータを初期化，その後，通常通り先行研究 [8] の手法を学習させた．

表 3 から，提案手法と先行研究を組み合わせた手法（Ours + Prev. [8]）は，元の手法からさらにスコアを向上させることがわかった．この結果から，提案手法はそれ単体としてだけでなく，モデルの初期化手法として汎用的に活用されうると考えられる．

(a)

Objects	man, footware, uniform
Prev. [8]	portrait of a happy young man in uniform
Ours	young man in a white uniform holds a baseball bat
Ours + Prev. [8]	young man in a white uniform holds a baseball bat

(b)

Objects	elephant
Prev. [8]	elephant walking through the river in the savuti , kenya
Ours	a elephant in a elephants
Ours + Prev. [8]	a elephant in a elephants

(c)

Objects	dog
Prev. [8]	cute dog lying on the bed in the morning
Ours	a dog with a cat and a bear
Ours + Prev. [8]	a dog is sitting on a bed with a cat

図 2　出力例．Objects は画像から検出された物体．

4.5　出力例

図 2 にモデル別のキャプション出力例を示す．物体名以外の単語について，提案手法が前置詞や冠詞など物体との共起頻度の高い単語を出力する傾向があるのに対して，先行研究 [8] は，“portrait of”，“in the savuti, kenya”，“in the morning” など物体との共起頻度が比較的低い記述を生成する傾向が見られる．このような共起頻度の低い記述は，いずれの例においても画像と一致していない．反対に，物体名については，(b) と (c) のように提案手法が余計な出力をする傾向がある．(c) では，先行研究の手法を組み合わせることでこの物体名の誤りが改善されている．

5　おわりに

本研究では，疑似教師ありキャプション生成において擬似キャプションが画像に対して部分的に一致しない問題に着目し，この齟齬のある部分を画像に対して教師信号として与えないよう抑制する手法を提案した．実験の結果，提案手法が先行研究の性能を上回り，疑似教師ありキャプション生成において上記の問題を考慮することの重要性が示された．また，提案手法は，モデルの初期化手法として先行研究の手法と組み合わせることで，さらに性能を向上させることが確認された．

参考文献

[1] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[4] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[6] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019.

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[8] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *CVPR*, 2019.

[9] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*, 2019.

[10] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, 2017.

[11] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. In *NeurIPS*, 2018.

[12] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018.

[13] Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. Compositional generalization in image captioning. In *CoNLL*, 2019.

[14] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *ICCV*, 2019.

[15] Fenglin Liu, Meng Gao, Tianhao Zhang, and Yuexian Zou. Exploring semantic relationships for image captioning without parallel data. In *ICDM*, 2019.

[16] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[18] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[19] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *the ninth workshop on statistical machine translation*, 2014.

[20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[21] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[22] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

A 実験設定詳細

	Training Text	Object Detector	Image Encoder	Text Decoder
[8]	Shutterstock	Faster-RCNN trained on OpneImages-v2	Inception-v4	1-layer LSTM of 512 dimension
[9]	Conceptual Captions	Faster-RCNN trained on OpneImages-v4	ResNet-101	1-layer GRU of 200 dimension

表 4 先行研究 [8, 9] それぞれの詳細な実験設定.

実験設定差分 4.1 節で述べたとおり，先行研究 [8, 9] ではそれぞれやや異なった実験設定をとっている．表 4 にその差分をまとめた．各要素の詳細は先行研究 [8, 9] を参照されたい.

疑似キャプションの前処理 画像から検出された物体名のセットに対して，2 つ以上検出された場合はそのうちのペアの組み合わせを用意し，各ペアが含まれている文を収集した．このとき，物体名の複数形の辞書（先行研究 [8]）を利用して，目的の物体名が複数形で含まれている文も取得する．収集した文から，目的の物体同士が関連した内容のものを選別するため，文中で目的の物体ペアの間に 4 つ以下の単語しか含まれない文を取り出した．また，目的の物体名ペアが複合語として出現している文を避けるため，物体名が隣り合っている文は取り除いた．同様に，各単体の物体についても文を収集した．ここで収集された文に対しては，目的の物体についてより詳細な記述のある文を選別するため，目的の物体とそれを修飾する形容詞の間に 2 つ以下の単語しか含まれない文を取り出した．統語解析器には spaCy [22] の en_core_web_lg を用いた.

学習イテレーション 頻出の物体に過適合することを防ぐため，各画像ではなく各物体ペアでイテレーションを行った．各物体ペアについて，これを含む画像と擬似キャプションをサンプリングして学習に用いる．また，物体ペア中の各物体についても同様にサンプリングと学習を行った．バッチサイズは 8 とし，validation セットでの CIDEr スコアが 20 エポック連続で更新されなくなった時点で学習を止めた．最適化には，Adam [23] を論文の推奨パラメータで用いた.

α の値 式 (7) について，$\alpha = 16$ とした．上述のように，学習は物体ペアとそのうちの各単体の物体を単位として行われるため，サンプリングされた文の中で検出された物体に対応する単語は 2 つか 1 つしかない．これに対してテキストデータの平均文長は，先行研究 [8, 9] でそれぞれ 12.0 と 10.7 である．この検出物体数と文長の比率を考慮して 2 の累乗数から最も性能が高くなる α を探した結果，文長 / 検出物体数 (= 1)，に近い，$\alpha = 16$ が最も高い性能を示した.

先行研究モデル初期化のための疑似キャプションの前処理 4.4 節の手法では，学習データの画像に対する提案手法の出力を疑似キャプションとして扱う．このとき，画像に対して全く異なる内容の疑似キャプションを与えることを防ぐため，次の前処理を行った．画像から検出された物体が 2 つ以上ある場合は，2 つ以上それらの物体を含む疑似キャプションのみ，検出された物体が 1 つの場合は，それを含む疑似キャプションのみを選別した．この基準を満たさない疑似キャプションについては，画像とともに先行研究モデル初期化のための学習データから外した.

B Ablation Study

		gate	*filter*	*unique*	*image*	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
[8]	Ours (full)	✓	✓	✓	✓	**49.5**	**27.3**	**13.1**	6.3	14.0	34.5	**31.9**	**8.6**
	w/o *filter*	✓		✓	✓	0.0	0.0	0.0	0.0	0.0	0.5	0.9	0.3
	w/o *gate*			✓	✓	40.9	21.5	10.1	4.8	12.7	32.1	17.6	6.0
	w/o *unique*	✓	✓		✓	47.2	26.2	13.0	**6.4**	**14.1**	**34.9**	28.3	8.5
	w/o *image*	✓	✓	✓		43.3	23.3	10.8	5.1	13.1	31.7	25.5	7.8
[9]	Ours (full)	✓	✓	✓	✓	**50.4**	**29.5**	**14.4**	**7.6**	**13.5**	**37.3**	**31.8**	**8.4**
	w/o *filter*	✓		✓	✓	44.5	25.4	12.2	6.2	12.4	36.7	29.2	7.5
	w/o *gate*			✓	✓	44.5	24.2	12.0	6.2	11.6	34.2	19.4	5.8
	w/o *unique*	✓	✓		✓	47.9	27.1	13.0	6.4	12.6	36.3	26.9	7.4
	w/o *image*	✓	✓	✓		47.1	26.0	12.8	6.6	13.1	34.7	29.7	8.0

表 5 Ablation study. 二重線の上下は，それぞれ先行研究 [8, 9] の実験設定に対応する．提案手法のスコアは 5 回の試行の平均，その他は一回の試行のスコア.

表 5 に ablation study の結果を示す．w/o filter は L_g だけを最小化したモデル，w/o gate は式 (1) のモデル化で L_g だけを最小化したモデル，w/o unique は 3.4 節のデコード制約を用いないモデルである．なお，3.3 節の疑似教師信号 f はゲート機構に対して与えられるものなので，ゲート機構なしで疑似教師信号だけ与えるモデル（w/o gate w/ filter）は比較しない．w/o image は，画像特徴量 $\boldsymbol{v}$ の代わりに，画像から検出された物体名の単語特徴量の平均を用いたモデルである.

いずれの実験設定においても w/o filter で特に数字が下がることから，3.3 節の疑似教師信号が性能の向上に大きく寄与していること，3.2 節のゲート機構はそれ単体ではうまく機能しないことがわかった．また，Ours (full) が突出してスコアが高いことから，提案手法はすべて組み合わせた状態で最も効果を発揮することが確認できる．w/o image のスコアが低いことについては，出力に物体検出器の誤りが伝播したものが多く見られており，これが原因だと考えられる.