強化学習における画像キャプションの低識別性問題と Long-Tail 分類手法を用いた対処

本多 右京 [1,2]　渡辺 太郎 [1]　松本 裕治 [2]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所
{honda.ukyo.hn6,taro}@is.naist.jp　yuji.matsumoto@riken.jp

概要

画像キャプションは，他の画像から入力画像を識別できるような，各画像に特徴的な情報を述べていることが望ましい．しかし，強化学習を用いたキャプション生成モデルは，高性能でありながら過度に一般的な内容のキャプションを生成してしまう．興味深いことに，強化学習はキャプション中の語彙を減少させる．本研究ではこの語彙の減少が低識別性の原因であると仮説を立て，long-tail 分類の手法を応用して既存モデルの語彙の増加を図る．実験の結果，提案手法が既存モデルの語彙を増加させるとともに，識別性も顕著に向上させることを確認した．また，低計算コストながら，識別性向上を図った先行研究に対しても識別性で上回ることを確認した．

1　はじめに

画像キャプション生成[1)]は画像の説明文を生成するタスクである．生成されたキャプションは，視覚障害者の補助 [1]，画像や動画の内容に基づく質問応答 [2, 3]，画像に関する対話生成 [4]，画像を用いたニュース生成 [5] など，様々な用途で活用されうる．

これらの用途で言及される情報は，各画像の特徴的な情報である．しかし，キャプション生成モデルは過度に一般的な内容のキャプションを生成してしまう（図 1 参照）．**識別性のあるキャプション生成**は，各画像の特徴的な情報を他の画像から入力画像を識別できる情報と捉え，この識別性を高めることを目標とする [6]．先行研究では，識別性向上に特化した新たな報酬やモデルが提案されてきた[2)]．しかし，これらの手法は計算コストの増加を伴い，さらにモデルを一から学習するコストを生じさせる．

本研究では，これらの計算コストを避け，**学習済みの既存モデルの識別性を高める**方法を検討する．

1）　以下，キャプション生成と表記．
2）　関連研究の詳細は付録 A を参照．

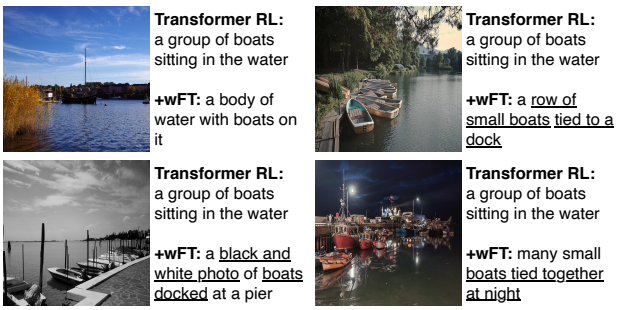

図 1　MS COCO 開発セットでの出力例．**Transformer RL** は強化学習を用いた Transfromer モデル，**+wFT** はそれに提案手法の fine-tuning を加えたモデルを指す．前者は同一の出力．後者はその他に特徴的な情報（下線部）を含む．

特に，識別性改善の余地が大きい，**強化学習を用いたモデル** [7] に着目する．キャプション生成での強化学習は，様々な評価指標での顕著な性能向上にもかかわらず，識別性に関しては改善がないか，あるいは低下させることが報告されている [8, 9]．興味深いことに，強化学習はキャプション中の語彙を減少させる [10]．モデルは実際に扱える語彙を超えた詳細を記述することが難しいため，語彙と識別性には強い関連がある．そこで，本研究ではこの語彙の減少が低識別性の原因であると仮説を立て，**識別性のあるキャプション生成を，既存モデルの語彙増加のタスクとして捉え直す**．語彙増加の目的関数は簡明であり，提案手法である，long-tail 分類手法を応用した少量の fine-tuning で達成可能となる．実験の結果，提案手法が既存モデルの語彙を増加させるとともに，識別性も顕著に向上させることを確認した．また，低計算コストながら，識別性のあるキャプション生成モデルに対しても識別性で上回った．

2　強化学習とその副作用

キャプション生成における強化学習　キャプション生成における強化学習は，微分不可能な評価指標スコアを直接最大化することを目的とする．REINFORCE [11] を適用することによって，評価指

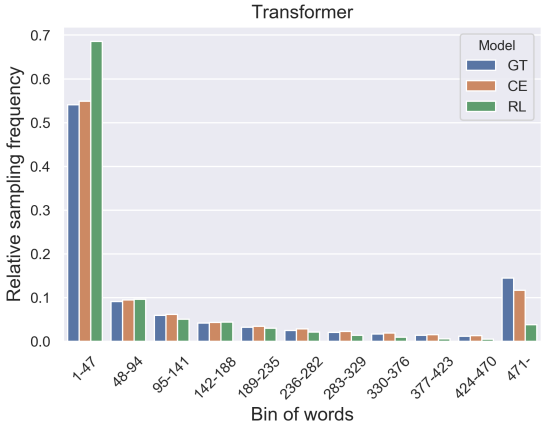

図 2　MS COCO の学習用の各画像に 5 文ずつサンプリングされた単語系列中の，単語の相対頻度．OOV を示す特殊トークン ⟨unk⟩ を除く 9,486 語を対象とした．単語は頻度順にソートされ，200 の bin に分割されている．最初の 10 個の bin と，残りの合計の相対頻度を表示．GT は正解キャプション．CE は最尤推定法，RL は強化学習で学習されたモデルの出力を示す．モデルは Transformer.

標スコアの微分を回避しながら，勾配を以下のように求めることができる [12, 7].

$$\nabla_\theta L_{\mathrm{RL}}(\theta) \approx -(r(w^s) - b)\nabla_\theta \log p_\theta(w^s \mid I). \quad (1)$$

ここで，$w^s = (w_1^s, \ldots, w_T^s) \sim p_\theta(w^s \mid I)$ は方策 p_θ からサンプリングされた単語系列，I は入力画像，$r(\cdot)$ は報酬関数，b は勾配の分散を安定させるための baseline 報酬である．特に Rennie ら [7] の強化学習手法は高性能で，キャプション生成モデルの学習におけるデファクト・スタンダードとなっている [13]. しかし，出力キャプションの語彙と識別性についてはむしろ減少させてしまうことがある [10, 8, 9].

強化学習による語彙の減少　Choshen ら [14] の研究は，強化学習と語彙の減少の関係を明らかにする．彼らは，強化学習がモデルの予測分布を peaky にすることを示した．強化学習では，方策 p_θ にもとづいて系列のサンプリングを行う．p_θ は，正解キャプションを生成する事前学習によって初期化される．ところが，テキスト生成モデル一般において，予測分布は高頻度語に偏ることが知られている [15, 16, 17, 18]．このため，強化学習では高頻度語のサンプリングと報酬の付与はできても，低頻度語についてはこれができない．この不均衡な報酬が，予測分布を高頻度語に向けてさらに peaky にし，出力する語彙をそれら高頻度語に限定させる．図 2 から，キャプション生成においても強化学習で高頻度語に peaky な分布が形成されることがわかる.

語彙の減少に伴う低識別性　このように，モデルが実際に生成できる語彙は高頻度語に偏る．もし画像中にモデルが実際に扱える語彙ではカバーできない詳細があった場合，モデルはその詳細を避け，高頻度語で記述できる情報だけを出力せざるをえない．そこで，本研究では，語彙の減少が強化学習モデルの低識別性の原因であると仮説を立てる.

3　提案手法

強化学習による語彙の減少は，予測分布が低頻度語から高頻度語に向けて偏ることによって引き起こされる．そこで，本研究では低頻度語の生成を促進することでこの語彙減少の問題に対処する．低頻度語の生成には，long-tail な不均衡データでの低頻度事例の分類手法が活用できる．以下では，これを応用した 2 つの fine-tuning 手法を提案する.

3.1　Simple Fine-Tuning

Simple fine-tuning (sFT) は，正解キャプション上での単純な fine-tuning である．これは，画像分類問題における long-tail 分類手法で強力なベースラインとなっている，Kang ら [19] の手法にもとづいている．彼らは分類器 $f_\theta(\cdot)$ を特徴抽出部分 $g_{\theta_e}(\cdot)$ と分類部分 $\boldsymbol{W}, \boldsymbol{b}$ に分解し[3)]，さらに学習も 2 段階に分解する．1 段階目の学習では，特徴抽出部分の学習を主な目的として，全データを用いて分類器の全パラメータを学習する．2 段階目の学習では，特徴抽出部分のパラメータは固定し，分類部分のパラメータのみを調整する．この調整の際，全データを使うのではなく，ラベルごとの事例数が均等になるようにサンプリングを行う.

これを，強化学習を用いたキャプション生成に応用する．1 段階目の特徴抽出部分の学習は，強化学習を用いた通常のキャプション生成学習に相当する．2 段階目で用いる，ラベルごとの事例数が均等なデータは，キャプションでは単語ごとの頻度の均衡がとれたデータに相当する．しかし，テキスト生成モデルからサンプリングされた系列は低頻度語を含まず，これを満たさない（2 節参照）．そこで，人手で作成された正解キャプションを，相対的に単語頻度の均衡がとれたデータとし，このデータの上で分類部分のパラメータの調整を行う.

学習データの語彙を $\mathcal{W}$ とし，ベクトル $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{W}|}$ の単語 w_i に対応する要素を $\boldsymbol{z}_{w_i}$ と表記する．また，ベクトル $\boldsymbol{z}$ に対して単語 w_i に対応する要素を返す

3)　$f_\theta(x) = \boldsymbol{W}^\top g_{\theta_e}(x) + \boldsymbol{b}$

softmax 関数を，$\Phi_{\beta,w_i}(\boldsymbol{z}) = \frac{\exp(\beta z_{w_i})}{\sum_{w_j \in \mathcal{W}} \exp(\beta z_{w_j})}$ とする．β は温度パラメータの逆数．これを用いて，入力画像 I の正解キャプション $w^g = (w_1^g, ..., w_T^g)$ の t 番目の単語 w_t^g の尤度は，以下のように計算される．

$$p_\theta(w_t^g \mid w_{<t}^g, I) = \Phi_{\beta, w_t^g}(s_\theta^t(w^g, I)), \tag{2}$$

$$s_\theta^t(w^g, I) = \boldsymbol{W}^\top g_{\theta_e}(w_{<t}^g, I) + \boldsymbol{b}. \tag{3}$$

特徴抽出部分 $g_{\theta_e}(\cdot)$ の出力の次元数を d とすると，$\boldsymbol{W} \in \mathbb{R}^{d \times |\mathcal{W}|}$，$\boldsymbol{b} \in \mathbb{R}^{|\mathcal{W}|}$．$g_{\theta_e}(\cdot)$ には Transformer [20] を用いた．分類部分の調整は，以下の **Cross-Entropy (CE)** 誤差を最小化する fine-tuning によって行う．

$$L_{\mathrm{CE}}(\hat{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\hat{\theta}}(w_t^g \mid w_{<t}^g, I). \tag{4}$$

パラメータ $\hat{\theta}$ は強化学習を用いて事前学習されている．パラメータの更新は，分類部分のパラメータ $\{\boldsymbol{W}, \boldsymbol{b}\} \in \hat{\theta}$ のみに対して，それぞれ勾配 $\nabla_{\boldsymbol{W}} L_{\mathrm{CE}}(\hat{\theta})$，$\nabla_{\boldsymbol{b}} L_{\mathrm{CE}}(\hat{\theta})$ を用いて行う．

3.2 Weighted Fine-Tuning

人手で作成された正解キャプションは，テキスト生成モデルからサンプリングされたキャプションに比べて低頻度語を含む割合が大きい．しかし，低頻度語は依然としてその頻度が低いことによって，高頻度語との間で学習の不均衡が残る．そこで，**weighted fine-tuning (wFT)** では，頻度への操作ではなく各単語の CE 誤差への重み付けを行うことによって，この不均衡をさらに是正する．強化学習モデルは，高頻度語に対しては高い確率を割り当て，低頻度語に対しては低い確率を割り当てるバイアスを学習している．**Bias Product (BP)** [21, 22] は，このようにバイアスを強く学習したモデルの確信度を利用することで，そのバイアスを取り除くように CE 誤差に重み付けを行うことができる．BP での w_t^g の尤度 $p_{\theta,\theta'}$ は以下のように計算される．

$$p_{\theta,\theta'}(w_t^g \mid w_{<t}^g, I) = \Phi_{w_t^g}\Big[\log \underbrace{\Phi_\beta(s_\theta^t(w^g, I))}_{p_\theta(\cdot \mid w_{<t}^g, I)} + \log \underbrace{\Phi_{\beta'}(s_{\theta'}^t(w^g, I))}_{p_{\theta'}(\cdot \mid w_{<t}^g, I)}\Big]. \tag{5}$$

ただし，$\Phi_\beta(\boldsymbol{z}) \in \mathbb{R}^{|\mathcal{W}|}$．パラメータ $\boldsymbol{\theta}$ **と** $\boldsymbol{\theta'}$ **は同じ事前学習モデルで初期化**されるが，θ は学習で更新されるのに対して $\boldsymbol{\theta'}$ **は固定**される．対数をとった上で p_θ と $p_{\theta'}$ を足し合わせ，再度正規化することで，p_θ は，事前学習モデルのバイアスを反映した $p_{\theta'}$ に対して相補的な値をとるよう学習される．付録 B に CE 誤差と BP 誤差を可視化した結果を示す．wFT の目的関数は，L_{CE} の p_θ を $p_{\theta,\theta'}$ に置き換えて，以下のように定義する．

$$L_{\mathrm{BP}}(\hat{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\hat{\theta},\hat{\theta}'}(w_t^g \mid w_{<t}^g, I). \tag{6}$$

sFT と同様に，パラメータ $\hat{\theta}$ と $\hat{\theta}'$ は強化学習によって初期化され，パラメータの更新は，分類部分のパラメータ $\{\boldsymbol{W}, \boldsymbol{b}\} \in \hat{\theta}$ のみに対して，それぞれ勾配 $\nabla_{\boldsymbol{W}} L_{\mathrm{BP}}(\hat{\theta})$，$\nabla_{\boldsymbol{b}} L_{\mathrm{BP}}(\hat{\theta})$ を用いて行う．

BP の先行研究に従い，**予測時は** $\boldsymbol{p_{\theta,\theta'}}$ **ではなく** $\boldsymbol{p_\theta}$ **だけを用いる** [21, 22]．これは，$p_{\theta'}$ の高頻度語へのバイアスを予測に取り込むことを防ぐためである．

4 実験

4.1 実験設定

データセットと評価指標　データセットには MS COCO [23, 24] の Karpathy 分割[4] [25]，評価指標には，BLEU-4 (BL-4) [26]，METEOR (MET) [27]，ROUGE-L (RG-L) [28]，CIDEr [29]，SPICE [30]，RefCLIP (RCLIP) [31] を使用した．先行研究 [8, 9, 32] に従い，識別性の評価指標には **R@K**[5] を用いた．

比較モデル　ベースライン（**Transformer RL**[6]）は，CIDEr スコアを報酬とした強化学習 [7] で訓練された，事前学習済みの配布モデルとした[7]．識別性のあるキャプション生成モデルとの比較には，R@K で最高性能を報告している **CIDErBtw** [9] と **NLI** [32] を用いた[8]．強化学習モデルの語彙を増やす手法として，**Joint CE** [10] との比較も行った[9]．また，CE 誤差最小化だけを強化学習モデルと同エポック数学習したモデル **Only CE** との比較も行った．

ハイパーパラメータの詳細は付録 C を参照．提案手法の各 fine-tuning は 1 エポックのみ行い，メモリ 16 GB の GPU1 枚を用いて約 10 分で完了した．

4)　頻度 5 以下の単語を ⟨unk⟩ に変換．語彙は 9,487 語．
5)　R@K は，出力キャプションを事前学習済みの画像–テキスト検索モデル [33] に入力した際に，他の画像の中から入力画像を上位 K 位以内で検索できたキャプションの割合．R@K が高いほど，出力キャプションの識別性が高いと判断される．
6)　LSTM [34] モデルでも実験したが，同様の結果のため割愛．
7)　https://github.com/ruotianluo/self-critical.pytorch `Transformer+self_critical` モデルを用いた．
8)　これらは，CIDEr 報酬に加えて識別性に関する報酬を強化学習で最大化する手法である．
9)　この手法では，L_{RL} と L_{CE} を同時に最適化することで，正解キャプション中の低頻度語のサンプリングを促す．しかし，依然として偏った分布からのサンプリングに依存することと，モデルを一から学習するコストが生じるという問題がある．

		Uniq-1	Uniq-S	Len	BL-4	MET	RG-L	CIDEr	SPICE	RCLIP	R@1	R@5	R@10
Transformer	**Transformer RL**	753	3,433	9.2	**39.0**	28.7	**58.7**	127.7	22.5	81.3	26.6	56.2	70.5
	+ sFT (Ours)	1,458	3,959	9.1	36.9	28.2	57.2	118.7	21.7	**81.5**	30.6	62.3	75.7
	+ wFT (Ours)	**1,776**	**4,274**	9.1	31.3	26.2	53.0	103.1	20.0	81.2	**32.5**	**64.5**	**77.1**
	CIDErBtw	837	3,609	9.5	38.6	28.8	58.6	128.2	22.6	81.2	27.7	57.6	71.6
	NLI	876	3,744	9.5	38.9	28.9	58.5	**129.1**	**23.0**	**81.5**	29.8	59.9	73.4
	Joint CE	1,083	3,491	9.3	38.6	**29.0**	58.3	123.8	21.9	81.2	27.3	57.2	70.8
	Only CE	935	3,599	9.4	35.0	27.7	56.0	112.2	20.8	80.9	26.5	55.8	69.7

表 1　ベースライン・先行研究との比較．*Uniq-1*・*Uniq-S* は，出力のうちの異なり語・異なり文の数．*Len* は出力の平均文長．提案手法の結果は灰色の背景色．モデルは全て Transformer．列の分割は左から，語彙，標準的評価，識別性に対応．

	識別性	正確性	流暢性
Transformer RL	3.00	4.42	4.83
+ wFT (Ours)	**3.34****	4.45	**4.84**
NLI	3.18**	**4.54**	4.76

表 2　人手評価の比較．ベースライン（Transformer RL）の識別性スコアは，相対評価の基準として 3.00 に固定されている．*/** は，ベースラインに対して，t 検定（識別性では 1 群 t 検定，それ以外では独立な 2 群の t 検定）で $p < 0.05/0.01$ での統計的有意差があることを示す．

4.2　実験結果

表 1 に，MS COCO テストセットでのベースライン・先行研究との比較結果を示す．

語彙　まず，sFT と wFT で Uniq-1 が顕著に増加していることから，提案手法が語彙の増加に成功していることがわかる．それにともなって Uniq-S も増加しており，各画像に対して個別のキャプションを生成できるようになっている．また，Len が増加していないことから，単に文長を長くしてこれらの効果を得ているわけではないことがわかる．sFT に比べて wFT の方が語彙の増加が大きいことから，BP による重み付けに効果があることがわかる．語彙増加を目的とした Joint CE や，強化学習を用いない Only CE もベースラインと比べて語彙の増加が顕著だが，提案手法が最も語彙の増加に成功している．

識別性　仮説のとおり，語彙の増加を直接の目的関数とした提案手法によって，識別性（R@K）が顕著に向上していることがわかる．さらに，識別性向上を図った先行研究である CIDErBtw と NLI と比較しても，非常に高い R@K スコアとなっている．これらの結果から，語彙の減少が強化学習モデルの識別性を低くする大きな要因であったことがわかる．

標準的評価　提案手法ではテキストベース指標（BL-4, MET, RG-L, CIDEr, SPICE）でベースラインからスコアが減少しているが，人間の評価との相関がより高いとされるテキスト・画像ベース指標の RCLIP [31, 35] では同等かそれ以上のスコアとなっている．このことから，提案手法の出力が質的に劣るわけではないことが示唆される．提案手法の出力には，正解キャプションには含まれないが正しい低頻度語が多く見られた．これが，正解キャプションにある情報しか評価できないテキストベース指標のスコアを実際より大きく下げていると考えられる．

4.3　人手評価

キャプションの識別性，正確性，流暢性について，Amazon Mechanical Turk を使って人手評価を行った．表 1 で高い性能を示したモデルを対象に，テストセットでの出力から 50 文をランダムに取り出し，1 文に対して 5 人の作業者に，5 段階で各基準のスコアを付与してもらった[10]．表 2 に結果を示す．識別性については提案手法が最も高いスコアを示し，ベースラインからも有意に高いスコアとなっている．正確性と流暢性についてはベースラインと同等であり，ベースラインと比較して質的に劣るわけではないことを示している．これは，4.2 節での RCLIP の結果と整合的である．

5　おわりに

本研究では，強化学習を用いたキャプション生成モデルの低識別性の原因が語彙の減少によるものと仮説を立て，long-tail 分類手法を応用して語彙を増加させる手法を提案した．実験の結果，仮説のとおり，提案手法が語彙の増加によって既存モデルの識別性を顕著に向上させることを確認した．強化学習モデルの低識別性の原因を明らかにし，低コストで実用的な識別性向上手法を提案する貢献と考える．

10）　ただし，識別性についてはスコアの絶対基準を設定することが難しいため，先行研究 [9] に従って**相対評価**とした．この相対評価では，ベースライン（Transformer RL）の出力と比較して特徴的な情報が述べられていれば最大 5，同じ情報なら 3，それ以下の情報なら最小 1 という基準を提示した．それ以外は，誤りの程度に応じて 5 から減点する絶対評価とした．

参考文献

[1] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In **ECCV**, 2020.

[2] Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan H Clark, and Regina Barzilay. Capwap: Captioning with a purpose. In **EMNLP**, 2020.

[3] Hyounghun Kim, Zineng Tang, and Mohit Bansal. Dense-caption matching and frame-selection gating for temporal localization in videoqa. In **ACL**, 2020.

[4] Julia White, Gabriel Poesia, Robert Hawkins, Dorsa Sadigh, and Noah Goodman. Open-domain clarification question generation without question examples. In **EMNLP**, 2021.

[5] Zhongping Zhang, Yiwen Gu, and Bryan A Plummer. Show and write: Entity-aware news generation with image information. **arXiv preprint arXiv:2112.05917**, 2021.

[6] Amir Sadovnik, Yi-I Chiu, Noah Snavely, Shimon Edelman, and Tsuhan Chen. Image description with a goal: Building efficient discriminating expressions for images. In **CVPR**, 2012.

[7] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In **CVPR**, 2017.

[8] Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. Generating diverse and descriptive image captions using visual paraphrases. In **ICCV**, 2019.

[9] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Compare and reweight: Distinctive image captioning using similar images sets. In **ECCV**, 2020.

[10] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In **CVPR**, 2019.

[11] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. **Machine learning**, 1992.

[12] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In **ICLR**, 2015.

[13] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on image captioning. **arXiv preprint arXiv:2107.06912**, 2021.

[14] Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. On the weaknesses of reinforcement learning for neural machine translation. In **ICLR**, 2020.

[15] Toan Q Nguyen and David Chiang. Improving lexical choice in neural machine translation. In **NAACL-HLT**, 2018.

[16] Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. On long-tailed phenomena in neural machine translation. In **Findings of ACL: EMNLP 2020**, 2020.

[17] David Demeter, Gregory Kimmel, and Doug Downey. Stolen probability: A structural weakness of neural language models. In **ACL**, 2020.

[18] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In **ICLR**, 2020.

[19] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In **ICLR**, 2020.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NeurIPS**, 2017.

[21] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In **EMNLP-IJCNLP**, 2019.

[22] He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. In **EMNLP-IJCNLP**, 2019.

[23] 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.

[24] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. **arXiv preprint arXiv:1504.00325**, 2015.

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

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

[27] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In **The Ninth Workshop on Statistical Machine Translation**, 2014.

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

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

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

[31] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In **EMNLP**, 2021.

[32] Zhan Shi, Hui Liu, and Xiaodan Zhu. Enhancing descriptive image captioning with natural language inference. In **ACL**, 2021.

[33] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In **BMVC**, 2018.

[34] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, 1997.

[35] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A Smith. Transparent human evaluation for image captioning. **arXiv preprint arXiv:2111.08940**, 2021.

[36] Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. **TACL**, 2017.

[37] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In **CVPR**, 2017.

[38] Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with character-level inference. In **NAACL-HLT**, 2018.

[39] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Group-based distinctive image captioning with memory attention. In **ACM MM**, 2021.

[40] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In **CVPR**, 2018.

[41] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In **ECCV**, 2018.

[42] Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. Joint optimization for cooperative image captioning. In **ICCV**, 2019.

[43] Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin. Fine-grained image captioning with global-local discriminative objective. **IEEE Transactions on Multimedia**, 2021.

[44] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In **NeurIPS**, 2017.

[45] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In **ICCV**, 2017.

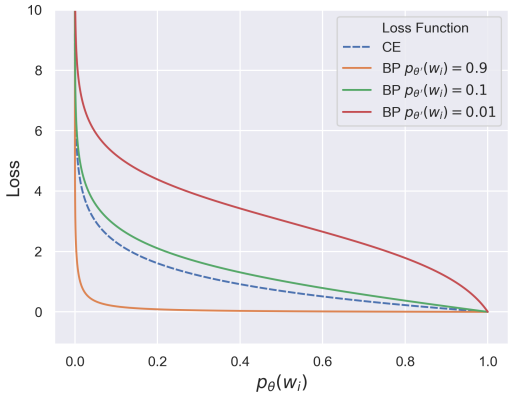

図 3　CE 誤差 $-\log p_\theta(w_i)$ と BP 誤差 $-\log p_{\theta,\theta'}(w_i)$ の可視化結果．BP 誤差の計算には $\{p_\theta(w_j)\}_{w_j \in \mathcal{W}}$ と $\{p_{\theta'}(w_j)\}_{w_j \in \mathcal{W}}$ の全単語の値を定める必要がある．ここでは，$i = 1$ として，$\frac{1}{5}(1 - p_\theta(w_1))$ を次の 5 つの要素 $w_2, ..., w_6$ に割り振った．これは，強化学習の予測分布では，最も確率の高い 5 つ単語が全体の 99%の確率を占めているという観測にもとづく．p_θ と $p_{\theta'}$ は同じパラメータで初期化されるため，上位 5 つの単語は両者の出力で同一と仮定し，$\frac{1}{5}(1 - p_{\theta'}(w_1))$ も次の 5 つの要素 $w_2, ..., w_6$ に割り振った．ここでは，$\beta = \beta' = 1$ とした．

A　関連研究詳細

識別性のあるキャプション生成の初期に主流であった手法は，入力画像に類似した画像を与えて，それらとは異なる情報を記述させる手法であった [30, 36, 37, 38, 39]．しかし，これらの手法では予測時にも類似画像を収集するコストがかかる．

そこで，予測時に類似画像を必要としない手法が提案されるようになった[11]．これらの手法は主に，識別性に関する報酬を新たに設計し，これを追加報酬として強化学習等で最大化するものである [40, 41, 42, 43, 9, 32]．その他には，負例のキャプションを用いて対照学習 [44] や敵対的学習 [45] を行うもの，正解キャプション[12]を単純なものと複雑なものに分割し，前者から後者への言い換えを行うモデルを提案したもの [8] がある．

上記のいずれの手法も，識別性を直接的に向上させる目的で，新たな報酬やモデルの設計を行っている．しかし，これらは計算コストの増加を伴い，さらにモデルを一から学習するコストを生じさせる．これに対して，提案手法は，既存モデルの語彙増加タスクとして捉え直されたタスクを扱う．これによって目的関数が簡明になり，既存モデルに対して少量の fine-tuning を適用するだけで識別性が向上する利点がある．

B　CE 誤差と BP 誤差の可視化

図 3 に CE 誤差と BP 誤差の可視化結果を示す．この図から，バイアスを反映したモデルの出力 $p_{\theta'}$ が単語 w_i について確信度が高いとき（$p_{\theta'}(w_i) = 0.9$）は BP 誤差は非常に小さくなり，逆に確信度が低いとき（$p_{\theta'}(w_i) \in \{0.1, 0.01\}$）には BP 誤差は CE 誤差と比較して大きくなることがわかる．

11）　本研究もこの設定に従っている．
12）　人手でアノテーションされたキャプションで，1 枚の画像に対して 5 つ付与されている [23, 24]．

Transformer RL: a tower with a clock on top of it
+wFT: a clock tower with a **weather vane** on top
NLI: a tower with a clock on the top of it
Human: a weather vane atop a cathedral clock tower

Transformer RL: a group of birds standing in the water
+wFT: a large group of **flamingos** stand in **shallow** water
NLI: a group of pink umbrellas are standing in the water
Human: a flock of pink flamingos standing in shallow water

Transformer RL: a black cat wearing a hat on top of a table
+wFT: a cat wears a **funny** hat while **staring straight ahead**
NLI: a black cat wearing a hat sitting on a table
Human: the cute black cat is wearing a bee's hat

Transformer RL: a dog next to a cup of coffee
+wFT: a dog is **sniffing** a cup of coffee
NLI: a dog standing next to a coffee cup on a table
Human: a squinting dog on a brick patio sniffs a cup of coffee

図 4　MS COCO 開発セットでの出力例．青字はベースライン（Transformer RL）の開発セット中の出力に一度も現れなかった単語を示す．ただし，正解キャプション（Human）は青字表記の対象外．

C　ハイパーパラメータ

ハイパーパラメータは，エポック数，学習率，式 (2,5) の β，式 (5) の β' 以外は全てベースラインモデルと同じものを用いた[13]．ただし，$\beta = 1$ としたので，β についてはベースラインモデルと同じである．提案手法の fine-tuning のエポック数はいずれにおいても 1 とし，学習率は{1e-3, 1e-4, 1e-5, 1e-6}，β' は{0.1, 1}のうち，開発セットでの R@1 スコアが最も高くなるものを選択した．ベストハイパーパラメータは，学習率は 1e-5，β' は 0.1 であった．モデルのパラメータサイズは，比較モデルも含めてすべて同一である．ただし，wFT での固定パラメータ θ' は，学習されず予測時にも使用されないので，このパラメータ数には含めていない．

D　出力例の分析

図 4 にキャプションの出力例を示す．提案手法出力（+wFT）では青字表記が多く，強化学習モデルの語彙に含まれない低頻度語が生成されていることがわかる．また，これらは各画像の特徴的な情報にもなっており，低頻度語の生成が識別性の向上に貢献していることがわかる．

13）　https://github.com/ruotianluo/self-critical.pytorch/blob/master/configs/transformer/transformer_scl.yml