文スタイル識別器による重みづけを用いた条件付き画像キャプション生成

松田　洋之[1]　大谷　まゆ[2]　脇本　宏平[2]
黒田　和矢[2]　中山　英樹[1]
[1] 東京大学大学院情報理工学系研究科
[2] 株式会社サイバーエージェント
{matsuda, nakayama}@nlab.ci.i.u-tokyo.ac.jp
{otani_mayu, wakimoto_kohei, kuroda_kazuya}@cyberagent.co.jp

概要

条件付き画像キャプション生成は，画像と条件を入力として，画像の説明文（キャプション）を生成するタスクである．本研究では，事前に学習したキャプション生成器の出力と，それとは別に学習した識別器の出力を混合して，所望のスタイル属性を持つキャプションを生成する．その際，識別器が出力する文スタイルの付与度に応じて，キャプション生成器の出力を適切に調整して混合を行う手法を提案する．

1　はじめに

条件付き画像キャプション生成は画像キャプション生成を拡張したものであり，画像に加えて，入力に新たな条件を追加する．このタスクでは，生成されるキャプションを条件により制御することを目標とする．

本研究では，文のスタイルを条件として入力に加える．画像キャプションデータセットを用いて，画像キャプションの文スタイルを制御する研究は広く行われている．しかし，先行研究の多くはスタイルを制御する機構がモデルの構造に依存しており，使用できるキャプション生成モデルが制限される．そのため，既存のキャプション生成モデルの活用が妨げられている．事前学習済みモデルを使用できる手法も存在するが，スタイルの制御機構を学習するために事前学習済みモデルの出力を必要とする．この場合，使用する事前学習済みモデルを変更すると，新たにスタイル制御の機構を学習する必要がある．

条件付き文生成の研究では，事前学習済みモデルへの依存度を下げたスタイル制御の手法（FUDGE）[1] が提案されている．この手法では，事前学習済みモデルを再学習せず，文スタイル識別器の情報を追加で利用する．そこで本研究では，この手法を画像キャプション生成に応用する．

事前学習済み画像キャプション生成器は，スタイル付きキャプションのデータによって学習されていないため，スタイル付きキャプションを構成する単語の出力確率を過小評価すると考えられる．本研究では，推論時に事前学習済みキャプション生成器の出力を変化させることによって，出力確率の過小評価を抑える機構を提案する．

2　関連研究

関連研究の一つに，条件付き文生成がある．CTRL [2] では，文のスタイルを表す Control Code を条件として多様な文の生成を可能にしているものの，データが大規模（140GB）なために学習コストが大きい．

近年では，学習コストを下げるため，事前学習済みモデルを再学習せずに生成文を制御する研究が行われている．PPLM [3] では，所望のスタイルを持つ単語が出力されやすくなるように，事前学習済み Transformer [4] の内部状態を更新している．GeDi [5]，FUDGE [1] では，事前学習済み言語モデルの出力を変化させ，所望のスタイルを持ちやすい単語の生成確率を引き上げることによって生成文を制御する．

条件付き画像キャプション生成の先行研究では，スタイル付き画像キャプションの公開データセット [6, 7]，が用いられている．Mathews ら [6] は，2 つの RNN を用意し，一方にスタイルの付かないキャプションを生成する役割を，もう一方にスタイル付き

キャプションを生成する役割を持たせている．Ganら [7] は 1 つの LSTM で生成を行うが，LSTM のパラメータをスタイルに共通の部分とスタイルに固有の部分に分けている．

MS COCO を代表とするスタイルの付かない大規模なデータセットに比べ，スタイル付きキャプションのデータ数は少ない[1)]．そのため，スタイル付き画像キャプションのペアデータを必要としない手法が提案されている [7, 9, 10, 11, 12].

3 手法

本研究では，条件付き文生成の先行研究（FUDGE [1]）の手法を用いて，事前学習済みモデルの単語の出力のうち，スタイルが付与されやすい単語の確率を高めるように補正する（3.1）．さらに，デコードの途中で，将来完成する文にスタイルが付与される可能性が低いと判断される場合，事前学習済みモデルの影響力を調整する手法を新たに提案する（3.2）．

3.1 文スタイル識別器による重みづけ

本研究では，画像キャプション生成に用いる言語モデルは自己回帰的であると仮定する．画像を I，スタイル条件 a を入力として，単語数 n のキャプション $X = (x_1, \ldots, x_n)$ を生成するスタイル付き画像キャプション生成器 $p(X|I, a)$ は，

$$p(X|I,a) = \prod_{t=1}^{n} p(x_t|I,a,x_{\leq t-1}) \tag{1}$$

と表される（$x_{\leq t-1} := (x_1, \ldots, x_{t-1})$）．ここで，$p(x_t|I,a,x_{\leq t-1})$ は，自己回帰的スタイル付き画像キャプション生成器である．ベイズの定理を用いると，

$$p(x_t|I,a,x_{\leq t-1}) = \frac{p(a|I,x_{\leq t-1},x_t)p(x_t|I,x_{\leq t-1})}{p(a|I,x_{\leq t-1})} \tag{2}$$

と変形できる．ここで，分子は x_t に依存しないから，

$$p(x_t|I,a,x_{\leq t-1}) \propto p(a|I,x_{\leq t-1},x_t)p(x_t|I,x_{\leq t-1}) \tag{3}$$

と表せる．$p(a|I,x_{\leq t-1},x_t) = p(a|I,x_{\leq t})$ は，時点 t までの出力単語と画像を入力として，将来完成されるキャプションがスタイル条件 a を満たすかどうかを識別する識別器である． $p(x_t|I,x_{\leq t-1})$ は，スタイルなし画像キャプション生成器である．

1） 例えば，MS COCO [8] にはスタイルのないキャプションが 1,026,459 含まれるのに対して，Senticap [6] に含まれる positive スタイルのキャプション数は 4,892 である．

式 3 より，スタイルなし画像キャプション生成器が学習済みのとき，スタイル付き画像キャプションを生成するためには，識別器のみを学習すればよい．本手法では，画像はキャプションスタイルに関与しないという仮定のもと識別器を近似し，時点 t までの生成文のみを入力して，スタイル（a）の付与度合いを予測する識別器 D_a を用いる．すなわち，

$$D_a(x_{\leq t}) := p(a|x_{\leq t}) \simeq p(a|I,x_{\leq t}) \tag{4}$$

とする．また，今後スタイルなし画像キャプション生成器を C で表すことにすると，式 3 は，

$$C(x_t|I,x_{\leq t-1}) := p(x_t|I,x_{\leq t-1}) \tag{5}$$

$$p(x_t|I,a,x_{\leq t-1}) \overset{\propto}{\sim} D_a(x_{\leq t})C(x_t|I,x_{\leq t-1}) \tag{6}$$

と表せる．

3.2 動的 Softmax Temperature

本研究では，式 6 における画像キャプション生成器 C に事前学習済みモデルを用いる．C はスタイルなしデータセットのみで学習されているため，スタイル付きのキャプションに用いられる単語の生成確率を極端に低く見積もる可能性がある．その場合，式 6 による生成において，識別器 D_a による補正の影響力が弱まり，スタイルが付与されづらくなると考えられる．

これを回避するためには，生成器 C の予測に温度パラメータ T を導入することで，C の出力を制御する手法が考えられる．温度パラメータは言語モデルの出力するロジットをスケール変換するハイパーパラメータであり，高いほど単語の生成確率の分布はなめらかになる．これにより，C の出力を制御し，識別器 D_a による補正の影響力を相対的に調整できる．その一方で，適切な温度パラメータは状況ごとに異なり，適切な値の設定は困難である．

そこで本研究では，温度パラメータを生成中の文のスタイルの付与度合いによって制御する機構を導入し，事前のパラメータ探索がなくても，適切な割合で生成器と識別器を混合する手法を提案する．

事前学習済み画像キャプション生成器 C が出力するロジットを C_{logit} で表す．温度パラメータ T を導入したときのキャプション生成器を C'_T で表すと，

$$C'_T(x_t|I,x_{\leq t-1}) := \text{softmax}_T\left\{C_{\text{logit}}(x_t|I,x_{\leq t-1})\right\} \tag{7}$$

$$:= \frac{\exp\left(\frac{C_{\text{logit}}(x_t|I,x_{\leq t-1})}{T}\right)}{\sum_{x'_t \in \mathcal{V}} \exp\left(\frac{C_{\text{logit}}(x'_t|I,x_{\leq t-1})}{T}\right)} \tag{8}$$

と表される．温度パラメーター T を適切に調整するために，T をスタイル識別器の出力確率の逆数 $1/D_a$ で置き換え，

$$C'_{1/D_a}(x_t|I, x_{\leq t-1}) \tag{9}$$

$$= \mathrm{softmax}_{1/D_a}\left\{C_{\mathrm{logit}}(x_t|I, x_{\leq t-1})\right\} \tag{10}$$

$$= \frac{\exp\left\{D_a(x_{\leq t-1})C_{\mathrm{logit}}(x_t|I, x_{\leq t-1})\right\}}{\sum_{x'_t \in \mathcal{V}} \exp\left\{D_a(x_{\leq t-1})C_{\mathrm{logit}}(x'_t|I, x_{\leq t-1})\right\}} \tag{11}$$

とし，

$$p(x_t|I, a, x_{\leq t-1}) \propto D_a(x_{\leq t})C'_{1/D_a}(x_t|I, x_{\leq t-1}) \tag{12}$$

によって単語 x_t を生成する．C_{1/D_a} 中で用いる D_a は時点 $t-1$ において計算されたものであることに注意されたい．

D_a は識別器の出力確率を表すから，$0 \leq D_a \leq 1$ である．スタイル付与度 $D_a(x_{\leq t-1})$ が 0 に近づくほど，C'_{1/D_a} は平坦な確率分布になる．その結果，式 12 において相対的にスタイル付与度 $D_a(x_{\leq t})$ の高い単語が生成されやすくなる．スタイル付与度 $D_a(x_{\leq t-1})$ が 1 に近づくほど，C'_{1/D_a} の分布は本来の確率分布 C に近づく．その結果，式 12 において相対的に $C'_{1/D_a}(x_t|I, x_{\leq t-1})$ が優先され，画像の内容に沿った単語が出力されやすくなる．

3.3 学習と推論

学習時には，スタイル付きキャプションとスタイルなしキャプションを用いて，文スタイル識別器 D_a を学習する．推論時には，式 12 に基づいてデコードを行う．本来であれば，$p(x_t|I, a, x_{\leq t-1})$ を得るために，モデル C の語彙 $\mathcal{V}$ 中の単語 $x_t \in \mathcal{V}$ すべてについて $D_a(x_{\leq t})$ を計算する必要がある．先行研究の FUDGE と同様，計算を効率的に行うために，$C(x_t|I, x_{\leq t-1})$ の出力確率が高い順に k 個を抽出し，抽出した単語についてのみ，$D_a(x_{\leq t})$ を計算する．それ以外の単語については，$D_a(x_{\leq t}) = 0$ であるとみなす．

4 実験

【実験設定: データセット】 SentiCap [6], FlickrStyle10K [7] を用いた．前者には positive, negative の，後者には humorous, romantic のスタイル付きキャプションが含まれている．SentiCap データセットは訓練データ，検証データ，テストデータに既に分かれている[2]ため，その分け方に従った．FlickrStyle10K データセットは，一部のデータ（訓練データ 7000 件）のみが公開されている．本研究では，この 7000 件のデータセットを再度訓練データ（72%），検証データ（18%），テストデータ（10%）にランダムに分割した．

【実験設定: モデル】 事前学習済み画像キャプション生成器として，CATR [3]を用いた．文スタイル識別器には，入力次元数 768，隠れ状態次元数 300 の 1 層 GRU を用いた．CATR は，トークナイズ手法として，huggingface transformers [13] の bert-base-uncased [4]相当のものを使用している．そのため GRU における単語埋め込みベクトルは bert-base-uncased の学習済みベクトルを利用した．

【実験設定: 手法】 ベースラインとして，事前学習済み画像キャプション生成器のみによる推論（original）を採用し，式 6 による推論（FUDGE），式 12 による推論手法（dynamic）と比較した．デコード方法には，条件付き文生成の先行研究で採用されている貪欲法を用いた．推論を効率的に行うため，文スタイル識別器 D_a に入力する単語は，キャプション生成器 C の出力確率の高い 200 単語に限った．

【実験設定: 学習・推論・評価】 訓練データを用いて，文スタイル識別器のみを学習する．検証データを用いて算出したクロスエントロピー損失が最も低いものを最良モデルとして選び，推論に利用した．評価指標としては，BLEU-1 (B-1), BLEU-3 (B-3) [14], METEOR (M) [15], ROUGE_L (R) [16], CIDEr (C) [17], SPICE (S) [18], discriminator score (dsc)，classification score (cls), perplexity (ppl), Self-BLEU-4 (SB-4) [19] を用いた．

4.1 実験結果

生成キャプションの定量的評価を表 1 に示す．4 データ全てにおいて，discriminator score, classification score の値が original, FUDGE に比べて改善している．このことから，提案手法によりスタイルが付与されやすくなっていることがわかる．

dynamic はスタイルが付与されやすいが，参照文との比較により評価を行う指標（BLEU，METEOR，ROUGE_L，CIDEr，SPICE）のほとんどでベースラインを下回った．提案手法において，生成器 C の影響が弱まると画像に沿ったキャプションが生成されにくくなることが主な原因であると思われる．ま

2） 訓練，検証，テストの順に，positive キャプションが 2380, 493, 2019 件，negative キャプションが 2039, 429, 1509 件となっている．

3） https://github.com/saahiluppal/catr

4） https://huggingface.co/bert-base-uncased

表 1　生成キャプションの評価．discriminator score は，文スタイル識別器 D_a の文末の出力の平均値である．classification accuracy では，文スタイル識別器 D_a とは別に，完成文のみを入力として文スタイルを持っているかどうかを判定する文スタイル識別器を，BERT に訓練データを用いて学習した．テストデータを 1 文ごとに識別器に入力し，出力確率が 0.5 以上となるものの割合を求めた．perplexity は，事前学習済みキャプション生成器 C によって計算した．

データ	手法	B-1 ↑	B-3 ↑	M ↑	R ↑	C ↑	S ↑	dsc ↑	cls ↑	ppl ↓	SB-4 ↓
positive	original	0.357	0.073	0.081	0.270	0.118	0.035	0.011	0.000	1.794	0.655
	FUDGE	0.365	0.074	0.086	0.271	0.113	0.036	0.058	0.022	1.944	0.662
	dynamic	0.315	0.057	**0.089**	0.251	0.085	0.033	**0.746**	**0.584**	4.272	**0.576**
negative	original	0.338	0.059	0.074	0.250	0.071	0.026	0.015	0.006	1.782	0.606
	FUDGE	0.357	0.062	0.079	0.254	0.078	0.028	0.052	0.030	1.891	0.608
	dynamic	0.305	0.050	0.076	0.233	0.055	0.024	**0.757**	**0.497**	5.012	**0.505**
humorous	original	0.242	0.066	0.104	0.250	0.372	0.150	0.019	0.001	1.866	0.744
	FUDGE	0.243	0.064	0.103	0.247	0.358	0.147	0.054	0.013	1.902	0.724
	dynamic	0.209	0.042	0.087	0.211	0.232	0.136	**0.539**	**0.326**	4.309	**0.665**
romantic	original	0.241	0.070	0.104	0.255	0.382	0.145	0.010	0.003	1.866	0.744
	FUDGE	0.263	0.072	0.106	0.257	0.375	0.142	0.030	0.009	1.963	0.726
	dynamic	0.224	0.047	0.090	0.229	0.254	0.126	**0.570**	**0.626**	3.472	**0.698**

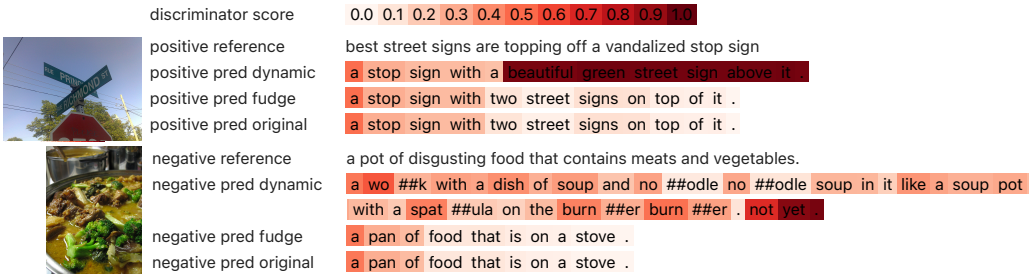

図 1　キャプションの生成結果の例．生成に成功した例と生成に失敗した例を一つずつ挙げる．デコード時にスタイル識別器 D が出力した確率を単語ごとに色を付けることで表しており，色と確率の対応関係は図中の discriminator score が表している．

た，スタイルの付与に大きく寄与する単語であっても，その単語が参照文に存在しない場合指標が改善しにくいことも原因の一つであると考えられる．

dynamic の手法を用いた場合，METEOR による評価は，positive データで original, FUDGE の手法を，negative データで original の手法を上回った．METEOR は参照文の類義語を考慮して評価を行うため，スタイルの付与に寄与する単語が評価されやすいことが理由として挙げられる．

perplexity は dynamic が最も高く，Self-BLEU は dynamic が最も低い．このことは，dynamic では生成されるキャプションが事前学習モデルにより生成されるものとは異なり，より多様になっていることを示唆している．

positive, negative のスタイル識別器を用いて生成した結果の一例を図 1 に示す．2 つの画像とも，FUDGE の生成文はスタイルの付与に失敗している．その一方で，dynamic は original, FUDGE とは異なる文を生成している．上の画像では，"beautiful"という positive な単語が出現しており，スタイルの付いたテキストを生成できている．一方で，下の画像のように，生成に失敗したものもある．dynamic の推論では，計算を効率的に行うために，キャプション生成器 C の出力確率上位 200 語のみを文スタイル識別器 D_a に入力している．この上位 200 語にスタイルの強い単語が含まれない場合であっても，この 200 語から単語を生成する必要があるため，スタイルづけの弱い単語が生成される可能性がある．そのような生成が続くと，画像の内容に沿わず，かつスタイルづけの弱い文が生成される．

5　おわりに

本研究では，事前学習済み画像キャプション生成器の出力と，文スタイル識別器の出力を混合して，所望のスタイル属性を持つキャプションを生成した．また，動的 Softmax Temperature を新たに提案し，スタイルが生成文に付与されやすくなることを確認した．生成文の主観的評価と，条件付きキャプション生成の先行研究との比較評価が今後の課題である．

参考文献

[1] Kevin Yang and Dan Klein. FUDGE: Controlled Text Generation With Future Discriminators. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3511–3535. Association for Computational Linguistics.

[2] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to Write with Cooperative Discriminators. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1638–1649. Association for Computational Linguistics.

[3] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In **International Conference on Learning Representations**.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc.

[5] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative Discriminator Guided Sequence Generation. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 4929–4952. Association for Computational Linguistics.

[6] Alexander Mathews, Lexing Xie, and Xuming He. SentiCap: Generating image descriptions with sentiments. In **Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence**, AAAI'16, pp. 3574–3580. AAAI Press.

[7] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. StyleNet: Generating Attractive Visual Captions with Styles. In **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 955–964.

[8] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, **Computer Vision– ECCV 2014**, Lecture Notes in Computer Science, pp. 740–755. Springer International Publishing.

[9] Alexander Mathews, Lexing Xie, and Xuming He. SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text. In **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 8591–8600.

[10] Cheng-Kuan Chen, Zhufeng Pan, Ming-Yu Liu, and Min Sun. Unsupervised Stylish Image Description Generation via Domain Layer Norm. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 33, pp. 8151–8158.

[11] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. MSCap: Multi-Style Image Captioning With Unpaired Stylized Text. In **2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 4199–4208.

[12] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. MemCap: Memorizing Style Knowledge for Image Captioning. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 12984–12992.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, prefix=von useprefix=true family=Platen, given=Patrick, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45. Association for Computational Linguistics.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318. Association for Computational Linguistics.

[15] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72. Association for Computational Linguistics.

[16] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out**, pp. 74–81. Association for Computational Linguistics.

[17] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 4566–4575.

[18] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, **Computer Vision – ECCV 2016**, pp. 382–398. Springer International Publishing.

[19] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**, SIGIR '18, pp. 1097–1100. Association for Computing Machinery.