クロスモーダル類似度に基づくカリキュラム学習による画像キャプション生成

張 宏寛 [1]　菅原 朔 [2]　相澤 彰子 [2]　周 雷 [1]　笹野 遼平 [1]　武田 浩一 [1]
[1] 名古屋大学　[2] 国立情報学研究所
{zhang.hongkuan.k5, zhou.lei.e1}@s.mail.nagoya-u.ac.jp
{saku, aizawa}@nii.ac.jp　{sasano, takedasu}@i.nagoya-u.ac.jp

概要

画像キャプションモデルには，様々な画像の内容をテキストで表現するための高度な汎化能力が求められる．多くの既存手法では，一般に画像とキャプションのペアの学習難易度の違いを考慮せず，モデルの訓練において同等に扱っている．一方，学習データの難易度を考慮するカリキュラム学習をキャプション生成に導入する手法が先行研究で提案されているが，難易度の測定には事前の定義や評価モデルの訓練が必要である．本稿では，シンプルかつ効率的な学習難易度測定法として，事前学習されたVision-Language モデルによって計算されたクロスモーダル類似度をキャプションデータの難易度として用いる手法を提案する．COCO および Flickr30k データセットを用いた実験の結果，提案手法は事前定義や追加の学習コストが不要で，ベースライン手法と同等以上の性能・収束速度を達成できることを示した．さらに，未知のデータやドメイン外のデータにおいて提案手法は高い性能を示し，汎化能力の向上が示唆された．

1　はじめに

画像キャプション生成は画像処理や自然言語処理分野で広く研究されている．しかし，多くの既存手法は，学習用の画像とキャプションのペアを区別せず扱っており，データごとの学習難易度の違いが無視されている．図 1 に示すように，キャプションデータセットでは，スタイルや複雑度が異なる参照文が一般に各画像に複数付与されている．この多様性はモデルの学習に偏りのある難易度を導入し，難易度の高いデータで訓練されたときにモデルを望ましくない方向に学習させる可能性がある [1]．

カリキュラム学習 (Curriculum Learning; CL) は，学習難易度に応じた順番でデータをモデルに学習させて，モデル性能と収束速度の向上を目指す手法である．しかし，CL を用いた画像キャプション生成の既存手法は，難易度測定に次のような欠点がある: 1) ドメインの専門知識やヒューリスティックが必要である点 [2]，2) 二つモーダルの難易度スコアを単純に加算しており，クロスモーダルな特徴を考慮できていない点 [3]，3) キャプションデータを用いて評価モデルを事前に訓練する必要がある点 (Bootstrap 手法) [2, 1]．

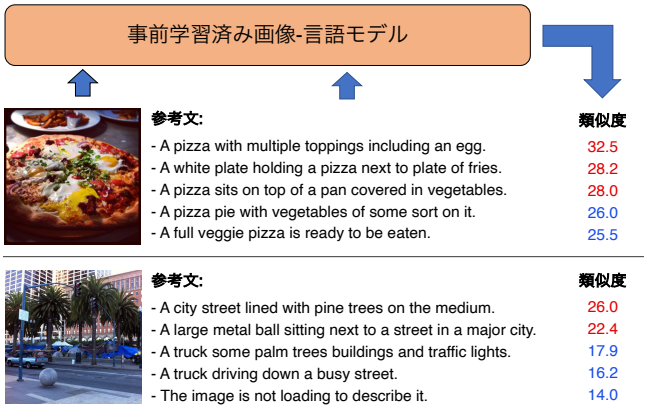

図 1　事前学習済み Vision-Language モデル CLIP により計算されたクロスモーダル類似度スコアの例．赤色と青色の数字は，それぞれスコアの高い方と低い方を示す．

本稿では，事前学習済み Vision-Language (VL) モデルを用いて，シンプルかつ効率的な難易度の測定手法を提案する．多くの VL モデルは，対応する画像とテキストをマッチングする事前学習タスクで訓練されるため，クロスモーダル類似度が計算可能である．この類似度は画像とテキストの関連性に対するモデルの確信度を表し，スコアが低いほど，判定が難しいか品質が低いデータであることを意味する [4, 5]．図 1 に示すように，シンプルな画像や適切な参照文を含むペアに高いスコアを与えて，逆に複雑な画像や関連性が弱い参照文を含むペアには低いス

コアを与える．本研究では，高いスコアを持つペアが関連性の高さから学習しやすいデータと考え，簡単なデータから順番にモデルを学習させる．

我々は，COCO [6] および Flickr30k [7] データセットを用いた実験の結果，クロスモーダル類似度に基づくカリキュラム学習で訓練されたモデルが，専門知識や追加の学習コストなしにモデル性能と収束速度の向上を達成することを示した．さらに，評価対象となるデータの知識を持つ VL モデルを用いるとさらに性能が向上した．また，未知のデータやドメイン外のデータにおいてより高い精度を達成し，汎化能力の向上を示した．最後に，提案手法をよりシンプルなモデルに適用することでも大きな性能改善が達成され，小さなモデルしか使えない場合でも提案手法が利用可能であることが示唆された．

2 関連研究

カリキュラム学習はモデルの汎化性能と収束速度を向上させる訓練手法として，機械翻訳 [8, 9] や画像分類 [10] などの研究に応用されている．データの学習難易度の測定は，事前定義を行う手法とモデルを利用する手法に分けられる．前者は画像中のオブジェクト数 [11] やテキストの長さ [12] などデータの特徴に基づくヒューリスティックにより測定する．後者はクロスエントロピー [13] やパープレキシティ [14] に基づくモデルの不確実性により測定する．

キャプションデータの難易度測定に関しては，Alsharid ら [3] は超音波画像を用いたキャプション生成において，視覚的難易度の測定に Wasserstein 距離を，言語的難易度の測定に TF-IDF を用いて両者を加算する手法を提案した．Dong ら [1] は複数の Bootstrap モデルを訓練し，各画像に複数のモデルでそれぞれ文生成を行い，生成文の BLEU スコアの平均値を難易度とした．Liu ら [2] は CT 画像のレポート生成において医学ドメインのヒューリスティックと Bootstrap モデル両方を用いて難易度を測定した．これらの研究とは異なり，本研究はクロスモーダル類似度を利用した効率的な難易度測定法を提案し，一般的なドメインのモデル性能の改善を目指す．

3 提案手法

3.1 クロスモーダル類似度

クロスモーダル類似度の計算には，ウェブ上の画像-テキストペアで事前訓練された CLIP [15]，および，人手でアノテーションされた画像-キャプションペアで事前訓練された ViLT [16] を用いる． CLIP を用いた類似度の計算では，P 個のパッチを含む画像 $X = (x_1, ..., x_P)$ と T 個のトークンからなるテキスト $Y = (y_1, ..., y_T)$ を与えると，CLIP はそれぞれのモーダルをエンコードして，視覚的な特徴 $\mathbf{x}$ とテキスト的な特徴 $\mathbf{y}$ を出力する．これらを用いて，類似度を式 (1) のように計算する．

$$D_{\text{CLIP\_sim}} = \cos(\mathbf{x}, \mathbf{y}) \tag{1}$$

一方で，ViLT を用いた類似度の計算では，[class] トークンを先頭とし，画像とテキストの入力を連結させ，クロスモーダルエンコーダによって式 (2) のようにエンコードする．

$$\text{ViLT}(X, Y) = x'_{[\text{class}]}, x'_1, ..., x'_P, y'_1, ..., y'_T \tag{2}$$

そして，この結合表現 $x'_{[\text{class}]}$ を，事前学習済み全結合層 FFN に与え，類似度を式 (3) のように計算する．

$$D_{\text{ViLT\_sim}} = \text{sigmoid}(\text{FFN}(x'_{[\text{class}]})) \tag{3}$$

3.2 訓練手法

クロスモーダル類似度によってソートされたデータでモデルを訓練するときには，データを与えるスケジュールを定める必要がある．本論文は，先行研究で広く使われている Baby Step [12] 学習スケジュールを用いる．具体的には，まずソートされたデータセットを L 個のサブセットに分割し，最も簡単なサブセットからモデルを訓練する．そして，検証セットでモデルの性能が数エポックにわたっても改善しない場合に，訓練が収束したとみなして，より難しいサブセットをマージして訓練を継続する．全てのサブセットをマージして収束を確認したら訓練を終了する．実験では，このスケジュールをすべての CL 手法に適用し，検証セットにおけるモデルの性能に基づいて最適なサブセットの数を調整する．この提案手法を Simi-CL と呼ぶ．

3.3 ベースライン

Addup-CL 各モダリティの難易度を計算して加算する．提案手法と公平に比較するために，各モダリティの難易度評価にも事前学習済みモデルを用いる．具体的には，視覚的な難易度 D_v を事前学習済み物体検出器 BUTD [17]，言語的な難易度 D_t を言語モデル GPT-2 [18] で測定し，両者のスコアを重み

付け和で加算して評価値 D_{addup} を得る．

$$
\begin{aligned}
D_v &= -\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n}\log p_{k,n},\\
D_t &= -\sum_{t=1}^{T}\log p(y_t|y_{<t}),\\
D_{\text{addup}} &= \lambda \times D_v + (1-\lambda)\times D_t.
\end{aligned}
\tag{4}
$$

ここで，K は画像から確信度が高い上位 K 個の検出ボックス，N は検出クラスの種類，$p_{k,n}$ は k 番目のボックスが n 番目のクラスに分類される確率，λ は重みを表す．

Bootstrap-CL　事前に学習および評価の対象となるデータセットでキャプションモデルを訓練して難易度スコアを算出する．具体的には，評価対象となるデータの訓練セットを用いて，一般的な訓練手法でデータの難易度を考慮せずにモデルを学習させ，訓練済みモデルで画像-キャプションペアに対して文生成のクロスエントロピーを計算する．

$$
D_{\text{bootstrap}} = -\sum_{t=1}^{T}\log p(y_t|y_{<t}, X) \tag{5}
$$

4　実験

4.1　実験設定

実験は COCO と Flickr30k データセットで行い，ともに Karparthy 分割 [19] を使用し，訓練/検証/テストセットのデータの割合はそれぞれ 113k/5k/5k と 29k/1k/1k とした．キャプションモデルは公開されているコード [20] に基づく Vanilla Transformer を実装し，ハイパーパラメータの設定は，バッチサイズを 10，学習率を 3e-4，ドロップアウト率を 0.4 とした．

CL 関連の設定について，Baby Step 学習のデータ分割数 L は，検証セットでの性能に基づいて COCO と Flickr30k でそれぞれ 5 と 3 とした．Addup-CL のハイパーパラメータは，物体検出クラスの数 N と検出ボックスの数 K をそれぞれ 1600 と 10 に設定し，加算の重み λ をパラメータチューニング後に 0.6 に設定した．Simi-CL の類似度計算に用いるモデルは，CLIP-base と ViLT-base 以外に，ViLT の作者から公開された評価対象となるデータ (COCO と Flickr30k) で微調整されたモデル (ViLT-CC と ViLT-FL と表記する) も用いて比較する．性能評価には BLEU-4 [21]，METEOR [22]，CIDEr[23]，SPICE [24] の 4 つの指標を用いる．

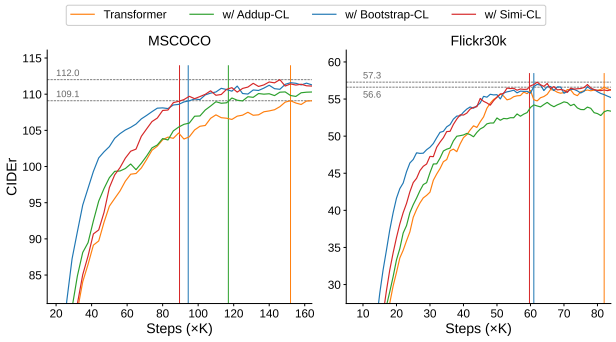

図 2　学習中の二つのデータセットの検証セットにおける Vanilla Transformer と CL ベース Transformer の性能曲線．

4.2　実験結果

検証セットでのモデルの性能を図 2 に示す．COCO では，全ての CL 手法が性能を向上させ，Vanilla Transformer の最高精度に達する収束時間が短くなっていることが分かる．特に，Simi-CL は Bootstrap-CL と比較して追加学習のコストなしに同等の性能を達成でき，また両方とも性能と収束速度が Addup-CL より優れている．Flickr30k でも同様の現象が観察できるが，改善の幅が COCO より小さく，Addup-CL は性能の改善が見られなかったため，CL 手法がより大きなデータセットに対して効果的であることが推測される．

テストセットでのモデルの性能を表 1 に示す，CL ベースモデルの性能は検証セットと類似している．Simi-CL のうち，ViLT-base を用いる場合は，Bootstrap-CL と同等の性能に達しているが，大量のデータから事前学習された CLIP-base の性能より下回っている．しかし，各評価対象となるデータセットで微調整された ViLT モデルを用いる場合は，ViLT-base や CLIP を上回る性能が達成されており，評価対象となるデータの知識を持つ教師モデルがカリキュラムの設計に役立つことが分かる．

4.3　考察

まず，カリキュラム学習されたモデルの汎化性能を考察するために，先行研究 [25] を参照して分布外データセットによる評価を行った．具体的には，COCO で訓練された最高性能のモデルを用いて，Flickr30k のテストセットでキャプションを生成して評価した，結果を表 2 に示す．Simi-CL によって学習されたモデルの性能が Vanilla モデルと Bootstrap-CL 手法を上回り，汎化能力が向上していることが分かる．

Model	COCO				Flickr30k			
	B@4	M	C	S	B@4	M	C	S
Our Implemented Baselines								
Transformer	35.7	27.9	113.0	20.9	27.7	21.8	58.5	16.0
Transformer + Addup-CL	35.2	27.9	114.2	21.0	26.5	21.5	56.6	16.0
Transformer + Bootstrap-CL	36.1	28.0	115.8	21.1	27.6	21.9	59.1	16.0
Our Proposed Methods								
Transformer + Simi-CL (ViLT)	35.9	28.0	115.6	21.2	27.3	21.9	59.0	16.0
Transformer + Simi-CL (CLIP)	36.3	28.1	116.2	21.2	27.0	**22.1**	59.6	16.2
Transformer + Simi-CL (ViLT-CC)	**36.4**	**28.2**	**117.1**	**21.4**	27.5	**22.1**	61.0	**16.3**
Transformer + Simi-CL (ViLT-FL)	36.0	28.0	115.9	21.0	**28.5**	**22.1**	**61.8**	16.2

表 1 COCO と Flickr30k におけるベースライン手法と提案手法で訓練されたモデルの性能の比較．B@4，M，C，S はそれぞれ BLEU-4，METEOR，CIDEr，SPICE を表す．

Model	B@4	M	C	S
Transformer	15.8	17.0	35.8	10.9
+ Bootstrap-CL	18.1	17.5	38.3	11.4
+ Simi-CL (ViLT-CC)	**18.6**	**18.2**	**39.8**	**11.7**

表 2 COCO における最高性能のモデルを用いて Flickr30k でテストした結果．

Model	in-domain		near-domain		out-domain	
	C	S	C	S	C	S
Transformer	69.3	10.9	59.5	10.0	35.5	6.8
+ Bootstrap-CL	70.5	10.9	61.9	10.4	35.5	7.1
+ Simi-CL (CLIP)	72.1	11.3	65.4	**10.7**	39.6	7.4
+ Simi-CL (ViLT-CC)	**72.8**	**11.4**	**65.5**	**10.7**	**40.2**	**7.7**

表 3 COCO における最高性能のモデルを用いて Nocaps の検証セットを評価した結果．

次に，カリキュラム学習されたモデルの困難なデータにおける性能を測定するため，Nocaps [26] データセットでモデルの性能を評価した．Nocaps では，COCO における画像のオブジェクトの種類と一致，部分一致および不一致の画像を含んでおり，その一致度に基づいてデータセットを in-domain，near-domain と out-domain というサブセットに分割した．Nocaps の設定に基づき，COCO で訓練された最高性能のモデルを用いて各サブセットでモデルの性能を測定した．表 3 に示される結果から，全てのサブセットで Simi-CL は Bootstrap-CL より高いスコアを達成し，特に困難な near-domain と out-domain サブセットで CIDEr スコアを大幅に向上させることが分かった．

最後に，Simi-CL が異なるアーキテクチャのモデルでも有効かどうかを調査するため，LSTM ベースのモデル BUTD と改良された Transformer モデル AoANet [27] に対しても Simi-CL を適用した．表 4 に示した結果のように，Simi-CL はよりシンプルな LSTM アーキテクチャに適用したときに改善の効果がより大きく，カリキュラム学習はサイズが小さいモデルでより効果的であることが示唆された．

Model	B@4	M	C	S
BUTD	35.2	27.2	109.9	20.1
+ Simi-CL (ViLT-CC)	**36.2**	**27.8**	**113.0**	**20.6**
AoANet	36.8	28.0	**117.2**	21.3
+ Simi-CL (ViLT-CC)	**37.3**	**28.2**	117.0	**21.4**

表 4 Simi-CL を異なるアーキテクチャのモデルに適用した場合の COCO における性能．

5 結論

本稿では，カリキュラム学習による画像キャプション生成のための，クロスモーダル類似度に基づく効率的な難易度測定手法を提案した．提案手法である Simi-CL はモデルの性能と収束速度を向上させ，特に大きなデータセットで効果的であった．また，評価対象となるデータセットの知識を持つ VL モデルを用いるとさらに性能が向上した．考察では，未知やドメイン外のデータで Simi-CL が最も高いスコアを達成し，モデルが高い汎化能力を得ることが分かった．最後に，Simi-CL を異なるアーキテクチャのモデルに適用し，よりシンプルなモデルでも効果が観察されたことから，小さなモデルしか実装できない場合にも応用できることが期待される．

参考文献

[1] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In **ACM-MM 2021**, pp. 2615–2624, 2021.

[2] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In **ACL-IJCNLP 2021**, pp. 3001–3012, 2021.

[3] Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. A course-focused dual curriculum for image captioning. In **ISBI 2021**, pp. 716–720, 2021.

[4] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. UMIC: An unreferenced metric for image captioning via contrastive learning. In **ACL-IJCNLP 2021**, pp. 220–226, 2021.

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In **EMNLP 2021**, pp. 7514–7528, 2021.

[6] 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.

[7] 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**, Vol. 2, pp. 67–78, 2014.

[8] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In **NAACL-HLT 2019**, pp. 1162–1172, 2019.

[9] Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. Norm-based curriculum learning for neural machine translation. In **ACL 2020**, pp. 427–436, 2020.

[10] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In **ECCV 2020**, pp. 247–263, 2020.

[11] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. **TPAMI 2016**, pp. 2314–2320, 2016.

[12] Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In **NAACL-HLT 2010**, pp. 751–759, 2010.

[13] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. Curriculum learning for natural language understanding. In **ACL 2020**, pp. 6095–6104, 2020.

[14] Yikai Zhou, Baosong Yang, Derek F Wong, Yu Wan, and Lidia S Chao. Uncertainty-aware curriculum learning for neural machine translation. In **ACL 2020**, pp. 6934–6944, 2020.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **ICML 2021**, pp. 8748–8763, 2021.

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In **ICML 2021**, pp. 5583–5594, 2021.

[17] 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**, pp. 6077–6086, 2018.

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, p. 9, 2019.

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In **CVPR 2015**, pp. 3128–3137, 2015.

[20] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In **CVPR 2018**, pp. 6964–6974, 2018.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL 2022**, pp. 311–318, 2002.

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.

[23] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In **CVPR 2015**, pp. 4566–4575, 2015.

[24] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In **ECCV 2016**, pp. 382–398. Springer, 2016.

[25] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In **CVPR 2011**, pp. 1521–1528, 2011.

[26] 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**, pp. 8948–8957, 2019.

[27] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In **ICCV 2019**, pp. 4634–4643, 2019.