空間的なラベル付けによるインタラクティブな画像キャプション生成

川辺 航 * 菅野 裕介 *

概要. 画像キャプション生成は，入力画像の内容を自然言語で出力するという画像認識タスクである．当タスクの特徴として，ある画像を表現しうるテキストの数は無数に存在し，その解が一通りに定まらないことが挙げられる．したがって，学習済みの機械学習モデルに推論をさせても，必ずしもユーザが望む出力が得られるとは限らない．しかし，モデルの挙動を更新するためには画像とそれに付随するテキストのペアを準備する必要があり，これは労力を要する作業である．この手間を減らすため，本研究では二次元空間上に画像とテキストを配置することで効率的にモデルの学習を実行することができるシステムを提案する．画像とテキストの視覚的かつ多対多の結びつきを可能にすることで，従来の単純なラベル付け手法と比較してユーザが入力すべきテキストの総量は減少する．本報告では，モデルの挙動が適切に更新されることを示すため，同一の画像群に対して異なるスタイルのテキスト群で学習を行い，推論結果を確認する．

1 はじめに

画像キャプション生成 [10, 8] は，入力画像の内容を自然言語で出力する画像認識タスクである．当タスクの難しさは，入力に対して出力が一意に定まらないことにある．すなわち，ある画像の内容を表現するテキストは無数に存在し，万人にとって満足のいく出力は不可能である．したがって，画像キャプション生成モデルがユーザにとって望ましい出力をするためには，ユーザ自身が自らの期待する出力を提示してモデルの挙動を更新する必要が生じる．例えば，特定の画像に対応するテキストをユーザが記述し，その画像とテキストのペアを用いてモデルを転移学習 [12] させるという方法が考えられる．しかしながら，画像とテキストのペアの作成にあたっては，数多くの画像一枚一枚に対してテキストをラベル付けする必要があり，手間と時間を要する．

本研究では，ラベル付けの労力を減らし学習の効率性を高めるために，GUI 上での視覚的な操作で画像テキストペアを作成し，モデルの挙動を更新することが可能なシステムを提案する．具体的には，画像とテキストの多対多の対応付け，テキストラベル自体の簡略化を通じて効率性を高める．本システムでは，図 1 に示すとおり，二次元平面上に画像やテキストを配置することで教師データ作成が完了する．このような視覚的なインタラクションを採用した理由として，空間配置によってユーザの認知的負荷が下がり [6, 7]，ラベル付けの効率性が上がる [1] ことが挙げられる．この空間において，一画像を複数テキストと紐づけることが可能であり，その逆もまた然りである．さらに，テキストは完全な文である必要がなく，文未満の単位であっても良い．具体的には，一画像に複数テキストが紐づく際，テキスト同士を結合することで擬似的な文を生成して教師データに含めることが可能である．これらの工夫よって，ユーザが記述するテキストの総量が減少し，モデル設計の効率性が高まる．本稿では実験として，特定のスタイルのキャプション群で学習を行い，モデルの挙動が適切に更新されることを確認する．

* 東京大学

2 システムの設計

画像テキスト変換モデルの挙動を変更する方法として，本研究では画像とテキストのペアに基づいてモデルの転移学習を行うことを考える．画像一枚一枚にテキストをつけていく方法は，(1) 画像の枚数分テキストを付与する必要があり，かつ (2) テキストを完全な文として記述する必要があるため，非効率的である．我々は，これらの課題を解消するべく，画像とテキストの多対多な紐付けが可能で，文未満の単位である単語やフレーズのラベル付けが可能な手法を提案する．

図 1 に示す通り，ユーザは画像と，範囲円付きのテキストを平面上にアップロードし，ドラッグ&ドロップでその配置を変更する．各テキストの範囲円の半径は変更可能であり，ある画像が円の内側にある時にのみ，その画像とテキストのペアは学習のための教師データに含まれる．こうすることで，一つの画像に複数のテキストを紐づけたり，一つのテキストに複数画像を割り当てることが可能になり，(1) が解決する．

また，テキストは必ずしも完全な文である必要はなく，単語やフレーズといった短い単位でもよい．ある画像に対しては複数のテキストが結び付けられているという状況が発生しうるが，その場合，テキスト同士を連結することで一つのテキストが新たに

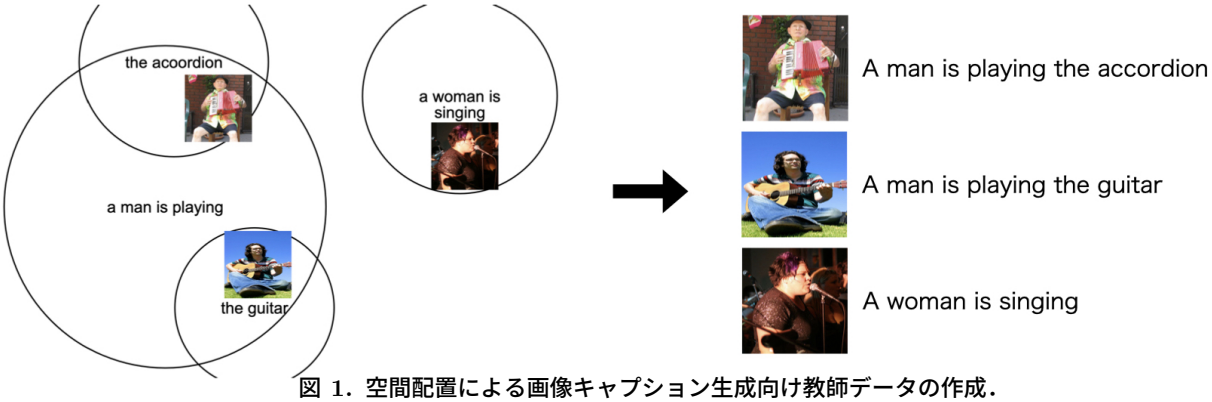

図 1. 空間配置による画像キャプション生成向け教師データの作成.

生成され，学習に利用される．テキスト同士の組み合わせによっては文法的な誤りが生じうるが，モデルは大規模コーパスで事前学習がなされているため，ある程度文法的な誤りを緩和して推論することが可能である．以上の工夫より（2）が解決する．

ユーザが学習実行ボタンを押すと，用意された画像テキストペアを用いて画像テキスト変換モデルの学習が行われる．学習後，アップロードされている全ての画像に対してモデルの推論結果が出力され，ユーザは画像上にマウスをホバーすることでその結果を確認することができる．ユーザは結果に応じて新たなテキストを生成したりテキストや画像の配置を変更することで，再度学習と推論のサイクルを回すことができる．

画像テキスト変換モデルとして，Transformer [9] ベースのモデル [3] を用いている．モデルはあらかじめ MSCOCO [5] および Visual Genome [4] で事前学習されており，学習の方法は元論文に則っているが，転移学習のエポック数は 50 に減らしている．なお，テキスト先頭部の Prefix は空である．また，キャプションの質を高めるため，学習時推論時ともに Beam search [2] を深さ 5 で使用している．さらに，一画像に複数テキストが紐づいている場合，テキストの範囲円がより大きい方（=より多くの画像に紐づけられると考えられる方）が先にくるように連結される．ただしこの部分に関しては改善の余地があり，既存の言語モデルによるソート，ユーザによる順序変更等を導入し検証する予定である．

3 実験

システムを用いてのモデルの学習の有効性を検証するため，本報告では三種類の異なるスタイルのキャプションを想定し，教師データの作成を行った上でモデルを訓練した．画像は Flickr30k [11] から 20 枚を抽出し，そのうち 5 枚に対し一画像につき一つのテキストを付与して教師データとし，学習後に残り 15 枚に対しての推論結果を出力した．結果が形式および内容の観点から我々の想定したものになっているかどうかは，著者の一人が判断した．

キャプションのスタイルとして，(パターン 1) 人物や動物の行動を名詞句（例：Swimming in the ocean, Walking on the road）で表現するケース，(パターン 2) 画像中の代表的なオブジェクトを 1 から 3 単語程度の名詞句で表現するケース（例：Dog, Baseball player），(パターン 3) 一般的なシーン記述的な文のケース（例：A man taking off his glasses is looking through a microscope），の三通りを想定した．

表 1. 学習後の推論に用いられた画像 15 枚のうち，形式や内容が我々の想定と一致していたものの枚数.

基準	パターン 1	パターン 2	パターン 3
形式	12 (80%)	15 (100%)	15 (100%)
内容	13 (87%)	14 (93%)	13 (87%)
両者	10 (67%)	15 (100%)	13 (87%)

学習後の推論結果に関する集計を表 1 に示す．教師データに使われる画像テキストペアは 5 という比較的少ない数であるが，モデルの出力がそれに応じて概ね適切に更新されていることがわかる．実際のユースケースでは 10 から 100 枚の画像へのラベル付けを想定しており，性能はさらに高まることが見込まれる．

4 おわりに

本稿では，GUI の操作を通じて画像キャプション生成モデルの挙動を更新することができるシステムを提案した．今後の方針として，一画像に複数テキストが割り当てられる際の教師データ中のキャプションの生成方法を定めたい．その上で，ナイーブなベースライン（例：画像一枚一枚に対して独立してテキストを割り振っていく UI）との比較実験を行い，使用感や学習後のモデルの性能を評価したい．

参考文献

[1] C.-M. Chang, C.-H. Lee, and T. Igarashi. Spatial labeling: leveraging spatial layout for improving label quality in non-expert image annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021.

[2] M. Freitag and Y. Al-Onaizan. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*, 2017.

[3] W. Jin, Y. Cheng, Y. Shen, W. Chen, and X. Ren. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021.

[4] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the IEEE European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

[6] T. W. Malone. How do people organize their desks? Implications for the design of office information systems. *ACM Transactions on Information Systems*, 1(1):99–112, 1983.

[7] F. M. Shipman III, C. C. Marshall, and T. P. Moran. Finding and using implicit structure in human-organized spatial layouts of information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 346–353, 1995.

[8] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara. From show to tell: a survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

[11] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[12] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.