視覚と言語の融合モデルにおける知識の振る舞いを調査するための表と画像の生成タスクの提案及びその調査結果

上垣外英剛 [1] 林克彦 [2] 渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] 北海道大学
{kamigaito.h,taro}@is.naist.jp katsuhiko-h@ist.hokudai.ac.jp

概要

本研究では視覚と言語の融合タスクである Vision & Language (V & L）において，自然言語から獲得されたエンティティに関する知識がどのように V & L モデルに保持されているかを検証するための新タスク，表と画像の生成を提案する．このタスクはエンティティと関連する画像からそれらに関する知識を含む表を生成するタスクと，エンティティとキャプション及び関連する知識を含む表から画像を生成するタスクの二つで構成される．我々は提案タスク遂行のためのデータセットを英語版 Wikipedia の Infobox から作成し，複数タスクで最高精度を達成している V & L モデル OFA で上記の検証を実施した．その結果，OFA はエンティティに関する知識の一部を事前学習時に忘却しており，それらの補完は良質な画像の生成に寄与することが判明した．

1 はじめに

近年，視覚と言語の融合タスクである Vision & Language (V & L) ではキャプション生成 [1] やテキストからの画像生成 [2] に代表されるような大きな成功を収めている．この進展の背景には大規模データにより事前学習された事前学習済み V & L モデルが存在している [3].

事前学習済み V & L モデルにおいて，入力に対して適切なキャプションや画像を生成するためには，事前学習済み V & L モデルが生成対象の特徴に関する知識を事前に保持している必要がある [4, 5]．現在，これらのモデルにおいて特にエンティティに関する知識は，自然言語処理で利用されている事前学習済み言語モデルのパラメータを引き継ぐことで，間接的に Wikipedia などのデータ資源を利用することで保持されている．このようにして V & L モデルに引き継がれた知識は，視覚と言語を横断したデータセットによって追加の学習を行うことで画像に関する表現と対応付けられる [6, 7, 8, 9, 10].

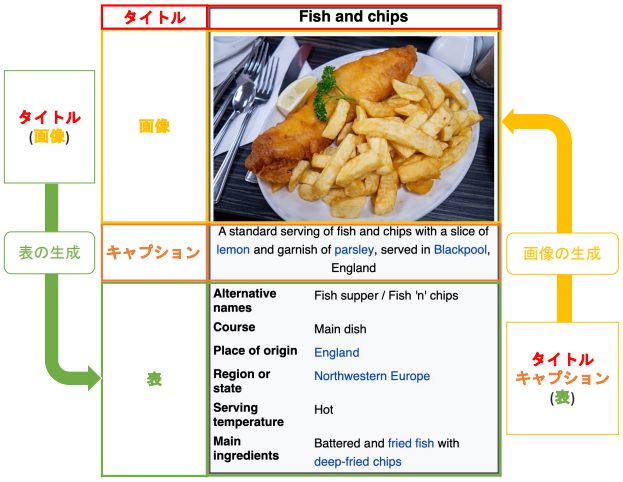

図 1 Wikipedia 記事中に含まれる Infobox の例[1)]．本研究では Infobox 中の画像及び表の箇所を生成することで，V & L モデルの検証を行う．

この学習過程において，自然言語から獲得された知識は事前学習済み V & L モデルの内部に適切に保持されているのか，あるいは画像の持つ特徴と組み合わせられることにより補強されているのかという問いは事前学習済み V & L モデルにより生成可能な対象の限界を知る上での重要な疑問である．

本研究ではそのような疑問を解消するために英語版 Wikipedia の各記事に含まれている Infobox を対象データとした，表と画像の生成タスクを提案する．図 1 に本研究が対象とする Infobox の例を示す．この例に示されているように，提案タスクでは，表または画像を生成する．いずれの場合においても適切に生成するためには，モデルが生成対象のエンティティに関する知識を把握していなければならない．

我々は提案タスクを遂行するために必要なデータセットを約 20 万件からなる Infobox を収集することで構築した．さらに我々は現在様々な V & L タスク

1） 引用: https://en.wikipedia.org/wiki/Fish_and_chips

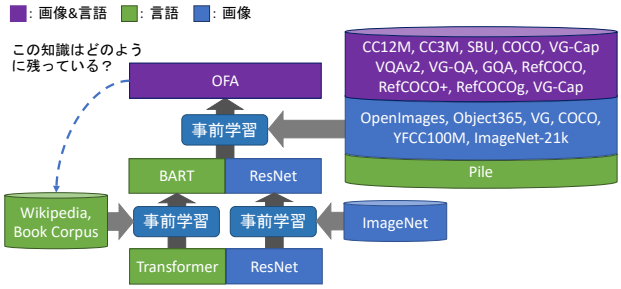

図 2 OFA の学習過程．本研究では事前学習時に Wikipedia から獲得されたエンティティに関する知識がどのように OFA に残されているのかを調査する．

で最高精度を達成している事前学習済み V & L モデルである OFA [10] を用いた検証を行った．

検証の結果，表の生成において，自然言語から獲得された知識の一部は V & L モデルにおける追加学習により消失することが判明した．また画像情報の利用により自然言語のみでは獲得されていなかった，エンティティが含む情報の種類に関する知識を新たに獲得していることが判明した．

画像生成においては，表に含まれる知識を利用することでより正確な画像を生成できることが明らかになった．また，自然言語だけで学習されたモデルにより推測された知識を利用することで生成される画像の多様性を向上させることが可能であることも判明した．

2 Vision & Language モデル

様々なタスクで最高精度を更新している事前学習済み V & L モデルの多く [6, 7, 8, 9, 10] は自然言語及び画像における事前学習済みモデルの重みを引き継いだ上で，自然言語と画像を横断したデータでの学習を行っている．本研究ではこのような学習過程を経て，自然言語を対象とした事前学習済みモデルに含まれていた知識がどのように変容するのかを検証する．なお，検証の対象としては複数の V & L タスクで最高精度を達成している OFA を選択する．

図 2 に OFA のネットワーク構造と各タスクとの関連を示す．[2] OFA ではデコーダ上で VQGAN [11] により画像を離散系列に変換して扱うことにより画像生成，自然言語生成を同一の Transformer [12] で行っている．また，この Transformer は BART [13] の重みを用いて初期化されるため，OFA には Wikipedia などの自然言語から獲得される知識が含まれていることが期待できる．なお，エンコーダではデコーダと異なり画像を直接扱うために，BART から引き継いだ埋め込み層に加え，ResNet [14] の出力を画像の埋め込みとして使用している．

2) 各学習過程で使用されているデータセットの詳細については付録 A に記す．

タスク	入力	出力
表の生成	タイトル, 画像	表
画像の生成	タイトル, キャプション	画像

表 1 各タスクの概要．それぞれの用語が指す Infobox 中の部位については図 1 を参照．

Alternative names | Fish supper / Fish 'n' chips<>Course | Main dish<>Place of origin | England<>Region or state | Northwestern Europe<>Serving temperature | Hot<>Main ingredients | Battered and fried fish with deep-fried chips

図 3 直列化された表の例．この例は図 1 の表を直列化したものである．

3 提案タスク: 表と画像の生成

本節では V & L モデルにおける知識の振る舞いを検証するための二つのタスク，**表の生成**と**画像の生成**についての説明を行う．いずれのタスクについても Wikipedia 記事中の Infobox に基づく．Infobox は Wikipedia 本文の情報に対応しているため[3]，事前学習済み V & L モデルが記憶している Wikipedia 中の知識を検証する上で適している．次節以降でそれぞれのタスクの詳細についての説明行う．

3.1 表の生成

表の生成タスクにおいて，対象となる V & L モデルは入力された Infobox のタイトルとしてのエンティティまたはそれに画像を結合した入力から表を生成する．表の生成に関しては説明文からの表の生成 [15] と同様に，表をテキストで直列化することで行う．我々の設定では，事前学習時の辞書に含まれるトークンを流用するために，列の区切り文字 | と行の区切り文字 <> を用いて図 3 のように直列化される．対象となるモデルの検証はこのような直列化されたテキストを直接生成することによって行われる．検証については次のような設定を用いて行う．

タイトルのみからの生成 事前学習済み V & L モデルと自然言語のみで学習された事前学習済みモデルにおいて，タイトルのみから生成した表を比較することで，V & L モデルが保持している自然言語のみを対象としたエンティティに関する知識を検証する．

3) https://en.wikipedia.org/wiki/Help:Infobox

画像を含む入力からの生成　画像を含む入力から表を生成し，その結果をタイトルのみを入力とした場合の結果と比較する．これにより，画像を考慮することで V & L モデルが新たに獲得した知識についてを検証する．

評価尺度　検証のための比較に際しては生成された表が実際のものにどれほど近いかを測ることにより行うため，次のような評価尺度を用いる．

ROUGE: 直列化された表はテキストデータであり，かつ Infobox は本文に対する要約の役割を担っていることから，自動要約の評価に使用されることが多い ROUGE [16] を使用する．なお，ROUGE での評価においては文字列の連なりが行や列に制限されないよう，列の区切り文字 | と行の区切り文字 <> をスペースに変換した上で行う．

Macro-F_1: 表の構造を考慮した評価を行うために，セルを種類毎に分けた上で参照となる表に対する一致を各事例毎に F_1 値によって評価し，平均化する．なお，一致を計算する際には，同一のセルが繰り返して出力されることによるスコアの増大を防ぐために，ROUGE の計算に使用されているクリッピングを適用する．セルの種類については下記のように分けて評価する．

- **グループ**: Infobox 中の表はグループに分けられることがあり，各グループの最初の行がグループ名を表すヘッダとなっている．グループ名に対する予測の精度は，モデルがエンティティに対しどのような側面での知識を持っているかを検証する上で重要である．
- **ヘッダ**: Infobox 中の 2 列以上の列からなる各行の先頭は通常，その同じ行で後続するセルのヘッダとなっている．従って，グループ名と同様の理由でヘッダへの予測精度は重要である．
- **値**: Infobox 中の 2 列以上の列からなる各行の 2 列め以降のセルはヘッダに対応する値を持っている．従って，値への予測精度はモデルがエンティティに対して詳細な知識を保持しているかを知る上で重要である．なお評価時にはヘッダとグループへの対応関係を考慮するために，対応するグループ名とヘッダを値と合わせた 3 つ組として扱う．

Micro-F_1: 上記の Macro-F_1 では事例毎に計算を行うため，モデルがどれだけ多様な知識を出力しているかを評価することが困難である．この問題を解決するために，事例全体を横断してセルを共有し，一括で F_1 値を計算する．

タスク	合計	訓練	開発	テスト
表の生成	204,850	184,480	10,098	10,272
画像の生成	86,862	78,201	4,271	4,390

表 2　各データセットのサイズ．

3.2　画像の生成

画像の生成タスクでは，モデルは Infobox のタイトルとキャプションまたは追加的に表を入力とし，対応する画像を生成する．検証については下記のような設定に基づいて実施する．

タイトルとキャプションからの画像の生成　画像生成に最低限必要な入力を用いることで，他のデータセットと比較した際の生成の難しさを検証する．

タイトルとキャプションと表からの画像の生成　表を含めた上で画像を生成し，その結果を上記の表を入力しない際の設定と比較することで，エンティティに関する知識が画像生成に与える影響について調査する．

評価尺度　画像生成の評価については現状で広く使われている下記の三つを使用する．

CLIP: 入力テキストと生成画像の関連度を測る尺度．事前学習済み V & L モデルである CLIP [17] によって推測される．

Inception Score (IS): それぞれの画像の差異を識別し易い，多様な画像が生成されていることを測る尺度 [18]．事前学習済み画像分類モデルである Inception-v3 [19] によって推測される．

Frechet Inception Distance (FID): 生成画像が参照画像にどれだけ近いかを測る尺度．IS と同様に Inception-v3 によって推測される．低いほど良い．

4　データセットの作成

データセットは英語版 Wikipedia の HTML ダンプデータ[4)]から Infobox を抽出することで作成した．生成対象の形式を揃えるため，抽出対象の Infobox は図 1 のように，一行目にタイトルを，二行目に画像を含むものに限定した．また，対象とする画像は jpeg 及び png 形式のものに限定した．なお，一部のキャプションについてはタイトルを含んでいないため，そのような事例に関してはハイフンを用いてキャプションの先頭にタイトルを結合した．

表 2 に各データセットのサイズを示す．なお，両タスクでデータセットサイズが異なるのは一部の

4)　https://dumps.wikimedia.org/other/static_html_dumps/current/en/

モデル	入力	ROUGE ↑			Macro-F ↑			Micro-F ↑		
		1	2	L	ヘッダ	グループ	値	ヘッダ	グループ	値
BART	Title	28.1	**13.3**	25.7	40.7	**22.8**	**2.4**	**61.5**	**34.3**	**5.5**
OFA	Title	27.6	13.2	25.3	38.4	21.2	2.3	59.2	33.9	5.3
OFA	Image	26.8	11.3	24.9	46.2	22.0	1.5	56.0	27.1	3.1
OFA	Title & Image	**28.6**	12.6	**26.3**	**46.8**	22.5	1.6	57.3	27.1	3.5

表 3　表の生成における各設定の結果．太字は最も高いスコアを，↑はスコアが高いほど良い結果であることを示す．

入力	CLIP ↑	IS ↑	FID ↓
Title & Caption	28.8	10.7	36.0
Title & Caption & Table (Gold)	**29.4**	**11.3**	**33.1**
Title & Caption & Table (BART)	28.0	10.8	37.6
Title & Caption & Table (OFA)	28.0	10.6	37.3

表 4　画像の生成における各設定の結果．↓はスコアが低いほど良い結果であることを示す．その他の表記については表 3 と同様．

Infobox がキャプションを含まないためである[5]．

5　検証

5.1　表の生成

設定　比較対象のモデルとして事前学習済み V & L モデルである OFA と自然言語のみで事前学習された BART を選択した．両モデルともに base モデルの重みを使用した．ハイパーパラメータについては OFA の論文で報告されているものを使用した[6]．

結果　表 3 に表の生成における各設定の結果を示す．タイトルのみを入力とした際には，BART の結果が OFA よりも高いことから，自然言語から獲得された知識の一部は V & L モデルにおける追加学習により消失することが分かる．また画像情報の利用によりヘッダに対する Macro-F_1 が改善されることから，画像を利用することでエンティティがどのような種類の特徴を持つかという知識が補強されることが分かる．その一方，セルの値に対する F_1 が改善されないことから，画像から得られる情報は自然言語から獲得された各ヘッダに対応する値のような詳細な知識を補うものではないことが分かる．

5.2　画像の生成

設定　表の生成と同様に，比較対象のモデルとして事前学習済み V & L モデルである OFA を選択した．表を入力する際には参照 (Gold)，§5.1 の出力結果のそれぞれを使用した．重みについては base モデルを使用した．ハイパーパラメータについては OFA の論文に記載されているものを使用した[7]．

結果　表 4 に画像の生成における各設定の結果を示す．この結果は CLIP の値では MS COCO [20] における画像生成の結果 [10] と近しいため，作成したデータセットのモデル学習における使用は妥当であると考えられる．また，表 (Gold) を入力することによりいずれの評価尺度も改善されていることから，エンティティに関する知識を補完してモデルに与えることにより，より質が高い画像が生成されることが分かる．この結果は OFA において Wikipedia に含まれるエンティティに関する知識が十分に保持されていないことを示している．

また，自動生成された表を与えた場合には CLIP 及び FID における性能の改善は確認できない．一方で BART により自動生成された表を使用することで IS については改善していることから，自動生成された表を用いることで出力画像の多様性が改善されることと，質が高い画像生成のためには表の予測精度が重要であることが分かる．

6　まとめ

本研究では事前学習済み V & L モデルに含まれている，自然言語のみから学習されたエンティティに関する知識が言語と画像を用いた追加学習を経てどのように保持されているかを検証した．

検証は英語版 Wikipedia から 20 万件の Infobox を抽出し，それらに含まれている画像と表を生成することで実施した．

事前学習済み V & L モデルである OFA を対象とした実験の結果，上記の知識は事前学習時に忘却されており，画像情報がその知識を完全に補っているわけではないことが示された．また画像生成においてはそれらの知識を適切に補うことで，より良質な画像が生成可能なことも示された．なお，本研究で使用したコードとデータセットは公開予定である．

5）　データセットの詳細は付録 B に記載．
6）　付録 C.1 に実験設定の詳細を記載
7）　付録 C.2 に実験設定の詳細を記載

謝辞

本研究は JSPS 科研費 JP21K17801 の助成を受けたものです.

参考文献

[1] 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 Francis Bach and David Blei, editors, **Proceedings of the 32nd International Conference on Machine Learning**, Vol. 37 of **Proceedings of Machine Learning Research**, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[2] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, **Proceedings of The 33rd International Conference on Machine Learning**, Vol. 48 of **Proceedings of Machine Learning Research**, pp. 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR.

[3] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. In Lud De Raedt, editor, **Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22**, pp. 5436–5443. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Survey Track.

[4] Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 4357–4366, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, **Computer Vision – ECCV 2020**, pp. 565–580, Cham, 2020. Springer International Publishing.

[6] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[7] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In **International Conference on Learning Representations**, 2020.

[8] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 07, pp. 11336–11344, Apr. 2020.

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 1931–1942. PMLR, 18–24 Jul 2021.

[10] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, **Proceedings of the 39th International Conference on Machine Learning**, Vol. 162 of **Proceedings of Machine Learning Research**, pp. 23318–23340. PMLR, 17–23 Jul 2022.

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2016.

[15] Xueqing Wu, Jiacheng Zhang, and Hang Li. Text-to-table: A new way of information extraction. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2518–2533, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[16] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 29. Curran Associates, Inc., 2016.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 2818–2826, 2016.

[20] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

[21] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2021.

表 5　OFA の事前学習に使用されているデータセットの一覧.

種類	タスク	データセット
言語&画像	Image Captioning Image-Text Matching	CC12M, CC3M, SBU, COCO, VG-Cap
	Visual Question Answering	VQAv2, VG-QA, GQA
	Visual Grounding Grounded Captioning	RefCOCO, RefCOCO+, RefCOCOg, VG-Cap
画像	Detection Image Infilling	OpenImages, Object365, VG, COCO OpenImages, YFCC100M, ImageNet-21K
言語	Masked Language Modeling	Pile

A　OFA で使用されているデータセットの詳細

OFA の事前学習では，言語，画像，言語と画像の各モダリティにおいて表 5 のように様々なデータセットが事前学習タスクのために使用されている．なお，表 5 に記載の Pile [21] には，英語版 Wikipedia の情報が 1.53%含まれている．従って，OFA の事前学習では V & L タスクに重きが置かれているものの，自然言語データから獲得された知識が忘却されないための工夫はなされていると理解できる．

B　作成したデータセットの詳細

Wikipedia HTML ダンプデータでは Wikipedia の各記事が HTML 形式で収録されているため，BeautifulSoup[8] を用いることで Infobox を抽出した．なお Infobox 中には [数字] といった形式で本文記事の参考文献へのリンクが含まれているため，それらは削除した．

表の生成において，入力用の画像については短辺が 480px を超える場合には縦横比を維持して短辺を 480px に縮小した．また，画像生成用の画像は縦横比を維持して元の画像の短辺を 256px に変更し，中央部を両辺 256px の正方形でクロップした．

なお，公開用データでは小規模なモデルから大規模なモデルまでの性能を測ることができるよう，表の生成では画像の短辺を 256px，384px までとしたデータセットも作成した．同様に画像生成では画像の両辺を 128px としたデータセットも作成した．

収集したデータの分割は，今後の拡張とデータの混同に配慮し，タイトルの SHA256 値を 20 で割った余りが 0 であればテストデータ，1 であれば開発データ，それ以外であれば訓練データとした．

データセットのサイズや抽出対象については本文を参照していただきたい．

8）https://www.crummy.com/software/BeautifulSoup/bs4/doc/

C　実験設定の詳細

両タスク共に OFA の実装には著者らが公開している実装[9] を改変して使用した．公開されている OFA では最大トークン長を決定する際にスペースで分割した後の単語数をトークン長としているため，BART などに合わせ，最大トークン長をサブワードで指定できるように変更した．なお，表の生成，画像の生成ともに我々は入力・出力共に最大長を 1024 サブワードに設定した．また，モデルとデータセットの特性を調査する観点から，学習は最尤推定のみで行い，強化学習については実施していない．

C.1　表の生成

BART と OFA の性能比較が実装の違いにより不公平とならないよう，BART の重みパラメータを OFA から引き継ぎ，OFA 上で BART を動作させた．学習時のハイパーパラメータはタイトルからの生成では OFA の自動要約における設定を引き継ぎ，画像を含む生成の際には OFA のキャプショニングにおける設定を引き継いだ．なお，公平な比較のために，推論時の設定は全てキャプショニングの設定に合わせた．実験は全て 4 枚の RTX 3090 で行った．

C.2　画像の生成

基本的には OFA で使用されているハイパーパラメータを引き継いだが，学習時間の関係で，各エポック終了後の開発データにおける画像生成時のビーム幅は 1 とした．なおテスト時は元の設定と同様ビーム幅 24 を使用している．表を使用する際には，入力の末尾に区切り文字 <> で結合して使用した．実験は全て 4 枚の RTX A6000 で行った．

9）https://github.com/OFA-Sys/OFA