教育指向 Text-と-Image ゲーム AI の検討

服部峻 [i]　高原まどか [ii]
[i] 滋賀県立大学 先端工学研究院　[ii] 龍谷大学 先端理工学部
hattori.s@e.usp.ac.jp,　takahara@rins.ryukoku.ac.jp

概要：近年、人工知能（AI; Artificial Intelligence）の研究分野では、自然言語文から画像生成する Text-to-Image や、逆に、画像から物体認識したり自然言語文を生成したりする Image-to-Text など、クロスメディア生成の研究が盛んであり、日進月歩している。一方、教育分野では、言語と画像（イメージ）とを脳内でクロスさせるトレーニングは、人間の高次な知性を支える心的過程の 1 つとも言われている。そこで本稿では、Text-と(to/from)-Image 技術を活用した教育指向ゲーム、及び、そのゲーム AI について検討する。今後は、プロトタイプを試作し、教育効果の検証を行っていく。
キーワード：Text-to-Image, Image-to-Text, 教育技術, ゲーム AI, クロスメディア

A Study on Education-oriented Text-to/from-Image Game AIs

Shun HATTORI[i]　Madoka TAKAHARA[ii]
[i] The University of Shiga Prefecture　[ii] Ryukoku University
hattori.s@e.usp.ac.jp,　takahara@rins.ryukoku.ac.jp

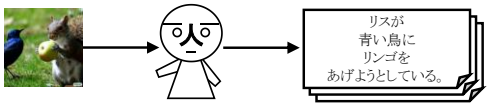
Abstract　In recent years, many researches on AI (Artificial Intelligence), especially, Text-to-Image, such as image generation from a natural-language text, and Image-to-Text, such as object recognition and natural-language text generation from an image, have been conducted actively and advanced constantly. Meanwhile, in the research filed of education, it is said that cross-media trainings between Texts and Images in a human's brain are one of mental processes to support his/her higher-level intelligence. Therefore, this paper conducts a study to build education-oriented text-to/from-image games and their game AIs. The future work plans to develop a prototype and validate its effects on education.
Keyword　Text-to-Image, Image -to-Text, EdTech, game AI, cross media

1. はじめに

近年、人工知能（AI; Artificial Intelligence）の研究分野では、自然言語文から画像を生成する Text-to-Image [1][2][3]や画像検索[4][5]、逆に、画像から物体認識したり自然言語文を生成したりする Image-to-Text [6][7][8][9]など、クロスメディア生成の研究が非常に盛んであり、日進月歩している。特に「敵対的生成ネットワーク（GANs; Generative Adversarial Networks）」による画像生成の進歩は著しく、layout-to-image [10][11][12]など，自然言語文による画像の条件指定だけでなく、画像内のオブジェクトの位置やサイズをレイアウト指定できるものもある。また、画像内のオブジェクト間のサイズバランスをよりリアルに調整する研究も進められている[13]。その他にも、対話応答などの Text-to-Text [14]や、類似画像検索などの Image-to-Image の研究も幅広く行われている。

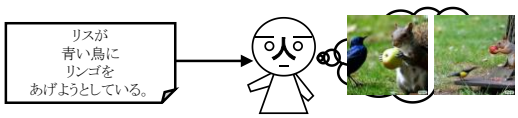
一方、教育分野では、言語と画像（イメージ）とを脳内でクロスさせるトレーニングは、人間の高次な知性を支える心的過程の 1 つとも言われ、着目されている[15]。第 1 に、図 1 のように、与えられた画像から、それを視覚で捉え、脳内で画像理解して、言語化する。画像内のオブジェクトを既知（物体認識できる）か否かといった事前知識、そのオブジェクトの色やサイズといった外観を表現する語彙力など、観測者に依存するため、出力（正解）は 1 つではない。入力である画像（表現）と比べて、出力であるテキスト（表現）の情報量は基本的に乏しいが、それでも多様性を有する。例えば、教育としては、『出来る限り他者にも解かり易く、画像を言語で表現し直せ。』といった問題が有り得るが、観測者の感性に依って、画像内オブジェクトの全てを言語化するのか絞るのか、取捨選択も伴う。観測者は自己本位ではなく、他者の（事前）知識や感性の多様性への理解や配慮にも繋がるかもしれない。第 2 に、図 2 のように、与えられたテキストから、それを視覚や聴覚で捉え、脳内で文章理解して、画像化（イメージ）する。出力である画像（表現）と比べて、入力であるテキスト（表現）の情報量は基本的に乏しいため、観測者の感性に依る補完が伴い、やはり多様性を有する。例えば、教育としては、『出来る限り他者にも解かり易く、テキストを画像（イラスト）で表現し直せ。』といった問題が有り得るが、前者の作文に比べると、作画行為自体の難易度が高く、一般向け、シングルプレイのゲームにするのは難しいかもしれないため、作画が好きな者を含む複数名で協力して楽しむゲームなどを検討する。

以上のように、本稿では、Text-と(to/from)-Image 技術を活用した教育指向ゲーム、及び、そのゲーム AI を検討する。今後は、プロトタイプを試作し、教育効果の検証を行っていく。



画像を視覚で見て、脳内で画像理解して、言語化する。
※入力に対して、出力（正解）は1つではなく、
人間の感性による取捨選択が伴い、それに依る多様性も有する。

図 1. 脳内での Image-to-Text トレーニング



テキストを読んで、脳内で文章理解して、脳内で画像化（イメージ）する。
※入力のテキストの方が情報量が乏しく、出力（正解）は1つではなく、
人間の感性による補完が伴い、それに依る多様性も有する。

図 2. 脳内での Text-to-Image トレーニング

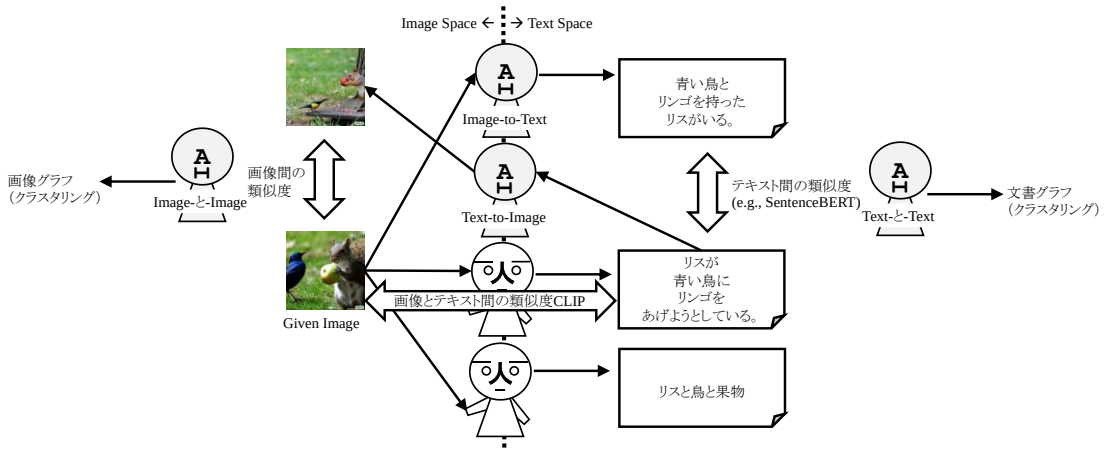


図 3. 画像とテキストとの間をクロスメディアする人間プレイヤーとゲーム AI

2. 教育指向 Text-と-Image ゲームとゲーム AI

近年、教育技術の分野においても「ゲーミフィケーション」は無視できず、いかに楽しく、主体的に学んでもらうかの工夫は重要である。本章では、Text-と(to/from)-Image 技術を活用した教育指向ゲーム、及び、そのゲーム AI を検討する。

従前、一般にも身近なものでは、シンプルなクイズやクロスワードパズルに始まり、テレビや YouTube などのクイズ番組においては、言語や画像を活用したゲームは多種多様に考案されて来ている。佐野ら[16]は、蓄積されたニュース番組から、画像付き選択クイズを自動生成する技術を提案している。

近年のゲームにおいて、言語や画像が全く活用されていないものはほぼ存在しないと思われるが、脳トレ要素が含まれていると考えられるものには、例えば、バンダイナムコエンターテインメントの『ことばのパズル もじぴったん』や、QuizKnock の『限界しりとり』という対戦ゲームなどが有名である。ナムコの『ワギャンランド』のボス戦では、複数のイラストが呈示され、そのイラストを用いて、ボス AI と互いに「しりとり」バトルするものもある。

2.1 シングルプレイ vs. マルチプレイ

脳トレ要素が含まれているゲームには、プレイヤー1 名がこつこつとシングルプレイするだけのものも無くはないが、少なくともゲーム AI との対戦プレイであったり、さらには、他のプレイヤーとの対戦プレイや協力プレイが活用されたりしている。

シングルプレイだけの場合、図 1 では、ゲームシステムはプレイヤーに画像を与え、プレイヤーはその画像を何らか言語化し、ゲームシステムはそのテキストを採点し、プレイヤーにフィードバックする。少なくとも、以下の機能が必要である。

・適当な画像をプレイヤーに与える
・画像に対して、プレイヤーのテキストを自動採点する

前者は、Web 上には膨大な画像が溢れており、著作権や Rating への配慮は必要であるが、技術的にはあまり問題無いと考える。一方、後者は技術的にも難しいが、openAI が開発した、画像とテキストとを繋げる汎用画像分類モデル CLIP [17]のスコアが活用できる。しかしながら、プレイヤーに依存しない汎用な技術であるため、1 章で前述した『出来る限り他者にも解かり易く、画像を言語で表現し直せ。』『出来る限り他者にも解かり易く、テキストを画像（イラスト）で表現し直せ。』など、多様性への理解や配慮の教育に繋げる事は出来ない。

逆に、図 2 では、ゲームシステムはプレイヤーにテキストを与え、プレイヤーは与えられたテキストを何らか画像化（イメージ）し、ゲームシステムはその画像を採点し、プレイヤーにフィードバックする。少なくとも、以下の機能が必要である。

・適当なテキストをプレイヤーに与える
・テキストに対して、プレイヤーの画像を自動採点する

後者は同様に CLIP を活用できるが、前者は必ずしも容易ではない。Web 上には膨大なテキストも溢れているが、プレイヤーの事前知識や語彙力と掛け離れたテキストでは、画像化（イメージ）するのが困難であり、ゲームとして不適当である。一般向けゲームにするにしても、個人特化させるにしても、どのようなテキストが適当なのか、十分に検討を深化させる必要がある。その上で、ゲーム AI がそれを適応的に選定する技術も考案する必要がある。

他のプレイヤーとの対戦プレイや協力プレイなどのマルチプレイの場合、プレイヤーに与える画像やテキストの選定プロセスにおいて、シングルプレイヤーだけでなく、マルチプレイヤーの事前知識や語彙力などを反映させる必要がある。自動採点プロセスでの活用を検討している CLIP 自体は、プレイヤーに依存しない技術であるため、シングルプレイでもマルチプレイでもそのまま適用できる。

最後に、ゲーム AI との対戦プレイや協力プレイなどのマルチプレイの実現には、少なくとも以下の機能も必要になる。

・与えられた画像から、人間らしく適当に言語化する
・与えられたテキストから、人間らしく適当に画像化する

最新の Text-と(to/from)-Image の技術を活用すれば、CLIP スコアで必ず満点を得るような「強い」ゲーム AI を構築できる可能性はあるが、強過ぎるゲーム AI との対戦プレイや、強過ぎるゲーム AI による協力は必ずしも楽しくなく、教育効果にも繋がらないと考える。ゲーム AI が対戦相手の人間プレイヤーに合った強さに適当に加減でき、かつ、人間らしく言語化したり、画像化したりできる必要もあり、難問である[18]。WordNet などの語概念階層データベースが活用できるかもしれない。語概念同士の上位・下位・同位関係、部分全体関係、特徴語など、自然言語処理の研究分野で盛んである。

2.2 主体的学習 vs. 育成ゲーム

これまで検討して来た教育指向 Text-と-Image ゲームでは、プレイヤーは、与えられた問題に解答する。そして、自身による Image-to-Text や Text-to-Image の出力（解答）に対して、ゲームシステムから高得点を得られるように成長するべく、オブジェクトに関する事前知識や、そのオブジェクトの色やサイズといった外観を表現する語彙力など、主体的な学習の促進が期待できると考える。

一方で、教育分野では、プレイヤーが問題を作成したり、あるいは、他のプレイヤーが問題を適切に解答できるように教え育てたりする事にも、教育効果が指摘されている。まず、前者に関しては、プレイヤーが作成した問題を他のプレイヤーやゲーム AI にも楽しんでもらうためには、2.1 節でのプレイヤーに与える画像やテキストの選定プロセスのように、問題を自動的にチェックする機能が必要になる。次に、後者に関しては、脳内、多種多様な事前知識レベル（テキスト、画像、及び、その繋がり）を有するようなゲーム AI を構築でき、かつ、プレイヤーによる教育という刺激によって、そのゲーム AI を適切に成長させる必要もあり、容易ではないが、語彙データベースを制御する事で簡易にシミュレートできるかもしれない。ゲーム AI を育成する脳トレゲームには、ゲームとしての新規性があり、かつ、新たな教育効果も期待できると考える。

2.3 ゲーム AI とのインタラクションの例

今後、教育指向 Text-と-Image ゲームのプロトタイプの要件分析や試作を進めて行くが、以下は、Image-to-Text ゲームにおけるプレイヤー（あなた）とゲーム AI との Discord インタラクションの理想とする一例である。但し、自動採点は CLIP スコアなどを活用しておらず、テキトーな数値である。

AI:「この写真って、何？」
あなた:「山」（3 点）
AI:「もうちょっと詳しく教えて？」
あなた:「大きな山」（5 点）
AI:「他には？」
あなた:「雪が降っている」（3 点）
AI:「この山はどこの？」
あなた:「。。。」（0 点）
AI:「日本一の」
あなた:「富士山？」（7 点）
AI:「そう！まとめると？」
あなた:「雪が降っている富士山」（10 点満点）

3. おわりに

本稿では、近年、人工知能の研究分野で日進月歩している、自然言語文から画像を生成する Text-to-Image や、逆に、画像から物体認識したり自然言語文を生成したりする Image-to-Text に着目し、それを教育分野における言語と画像（イメージ）とを脳内でクロスさせるトレーニングに応用できないかと考え、Text-と(to/from)-Image 技術を活用した教育指向ゲーム、及び、そのゲーム AI を検討した。今後は、プロトタイプを試作し、教育効果の検証を行っていく。

参 考 文 献

[1] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022) Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv:2204.06125.
[2] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022) Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, arXiv:2205.11487.
[3] Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z. R., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., and Wu, Y. (2022) Scaling Autoregressive Models for Content-Rich Text-to-Image Generation, arXiv:2206.10789.
[4] Hattori, S. (2012) Peculiar Image Retrieval by Cross-Language Web-extracted Appearance Descriptions, International Journal of Computer Information Systems and Industrial Management (IJCISIM), Vol.4, pp.486-495, MIR Labs.
[5] Hattori, S. (2013) Hyponymy-Based Peculiar Image Retrieval, International Journal of Computer Information Systems and Industrial Management (IJCISIM), Vol.5, pp.79-88, MIR Labs.
[6] 服部 峻, 手塚 太郎, 田中 克己（2007）文書中の地物画像を言語的記述で代替するための地物の外観情報の Web からの抽出, 情報処理学会論文誌（トランザクション）データベース, Vol.48, No.SIG11 (TOD34), pp.69-82.
[7] Vladimirov, L. (2020) TensorFlow 2 Object Detection API tutorial, https://tensorflow-object-detection-api-tutorial.readthedocs.io/en/latest/index.html
[8] Redmon, J., and Farhadi, A. (2018) YOLOv3: An Incremental Improvement, pp. 1-6, arxiv.org:1804.02767.
[9] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv:2004.10934.
[10] Sun, W., and Wu, T. (2021) Learning Layout and Style Reconfigurable GANs for Controllable Image Synthesis, IEEE Transactions on Pattern Analysis and Machine Intelligence, arXiv:2003.11571v2.
[11] He, S., Liao, W., Yang, M. Y., Yang, Y., Song, Y.-Z., Rosenhahn, B., and Xiang, T. (2021) Context-Aware Layout to Image Generation with Enhanced Object Appearance, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21), pp.15049-15058.
[12] Tang, H., and Sebe, N. (2021) Layout-to-Image Translation with Double Pooling Generative Adversarial Networks, IEEE Transactions on Image Processing, Vol.30, pp.7903-7913.
[13] 服部 峻, 相場 築, 高原 まどか, 工藤 康生（2022）画像内オブジェクトのサイズバランス調整支援システム, WebDB 夏のワークショップ 2022.
[14] 森 康汰, 服部 峻, 高原 まどか, 工藤 康生（2022）ツンデレな対話応答 AI の構築コスト削減のためのクロス言語とルールベース個性除去, WebDB 夏のワークショップ 2022.
[15] 北神 慎司（2003）画像の記憶における言語的符号化の影響に関する研究, 京都大学 博士論文（教育学）.
[16] 佐野 雅規, 八木 伸行, 片山 紀生, 佐藤 真一（2009）蓄積されたニュース番組からの画像付きクイズ生成手法の提案, 電子情報通信学会論文誌 D, Vol.J92-D, No.1, pp.141-152.
[17] openAI (2021) CLIP (Contrastive Language-Image Pre-Training), https://github.com/openai/CLIP.
[18] 服部 峻, 吉田 裕太, 高原 まどか（2021）ヒト型化オセロ AI を用いたビデオゲームのインタフェース改善, ヒューマンインタフェース学会論文誌, Vol.23, No.4, pp.459-480.