類音語の連想性を考慮した未知語の発音に対する画像生成

松平 茅隼[1,a)]　カストナー マークアウレル[2,b)]　駒水 孝裕[1,c)]　平山 高嗣[3,1,d)]　道満 恵介[4,1,e)]
井手 一郎[1,f)]

概要

テキストから画像を生成するモデルが近年急速に発展する一方で，定義が存在しない語（未知語）に対する生成画像が人間の直感と合致しないという問題点はあまり着目されてこなかった．未知語に対する生成画像が人間の直感と相違する原因として，人間が未知語の発音から直感的に意味や印象を想像，またその類音語を連想する心理学的現象がある．そこで本発表では，類音語の連想性を考慮した未知語―画像間の対応付けを行なうことで，未知語に対するより直感的な画像生成を実現する．主観評価実験により，提案手法がテキストから画像を生成する既存手法と比べ，未知語に対してより人間の直感に合致した画像を生成することを確認した．また，提案手法を多言語へ拡張しその画像生成結果を比較することで，他言語間で類似／相違した概念を連想する発音の探索を試みた．

1. はじめに

近年，テキストから画像を生成する大規模マルチモーダルモデルが注目されている．ここ数年間で，DALL·E [13], [14] や Stable Diffusion [15] に代表される様々なモデルが提案され，入力したテキストに合致した高品質な画像の生成が容易となった．しかし，それらが行なうテキスト―画像間の対応付けは必ずしも人間的でなく，特に入力したテキストに明確な定義が存在しない造語等の語（未知語）が含まれる場合，人間の期待に沿わない画像が生成される傾向が指摘されてきた [7], [9]（図 1(a) 参照）．

一方，人間は無意味に見える言語音の羅列に特定の概念や印象を関連付けることがあり，これは音象徴性と呼ばれる [4]．例えば，未知語「ブーバ」は未知語「キキ」よりも

[1] 名古屋大学
[2] 京都大学
[3] 人間環境大学
[4] 中京大学
[a)] matsuhirac@cs.is.i.nagoya-u.ac.jp
[b)] mkastner@i.kyoto-u.ac.jp
[c)] taka-coma@acm.org
[d)] t-hirayama@uhe.ac.jp
[e)] kdoman@sist.chukyo-u.ac.jp
[f)] ide@i.nagoya-u.ac.jp

$ 2: Š Z € p Š b ” ¨ ; w È Ý Q › ß € ` h ° Œ w C ; t 0 b ” h þ \ R O w “ A }

2. È Z €

b | Ú ™ w Â © µ Ä › Ú E q ` o h þ › \ R b ” G F Û Þ Ã ç w Z € U ] œ p K ” } VQGAN-CLIP [2] • DALL ·E [14] t • ‡ “ | f w ™ w ¦ „ Þ Ã ç [18] w Z q t ™ ! ` ^ • | GLIDE [10] | DALL ·E 2 [13]| Stable Di u-sion [15] s 7 ‘ s Þ Ã ç U Š ^ • h }

\ • ’ w X t ž è b ” › Ã x | Â © µ Ä r g æ ü › Ä ² ¶ 6 A ^ t Þ Ã ç t ‘ ` o M ” : p K ” } « Q y Stable Di usion p x | h þ \ R w Ú E Ç Z t Contrastive Language Image Pretraining (CLIP) [11] w Â © µ Ä ¤ ï ¯ ” ¼ w Z — U f w ‡ ‡ ; M ’ • ” } \ • t ‘ “ | ° Œ t 0 b ” • ^ w Æ † ^ › ‰ Ä ² ¶ 6 A ^ t Þ Ã ç U Ë m Q í U | h þ \ R A L t ‹ S é ^ • o ` ‡ O }

f \ p Š æ p x | Stable Di usion w CLIP › ; M h Â © µ Ä r g æ ü › ~ ! b ” \ q p | h þ \ R t S Z ” ° Œ • h þ w 0 Ç Z › M š b ” } CLIP w › Ã í t 0 ` o ý h t ; ` ¶ $ Œ _ › ‹ Ö b ” \ q p | C ; ¨ Å S › ‘ “ Y ¬ t ß € ` h ° Œ T ’ w h þ \ R › æ s O }

3. ° Œ w C ; t 0 b ” h þ \ R

$ 2 t Ô b ‘ O t | Š O x f ‘ U Œ æ Z € p Š ` h C ; ¤ ï ¯ ” ¼ IPA-CLIP [8] q h þ \ R Þ Ã ç Stable Di usion [15] q M O 2 ¨ w Þ Ã ç T ’ s ” } IPA-CLIP x | Ö — ^ • ” C ; › CLIP [11] Â © µ Ä ¤ ï ¯ ” ¼ q ‰ ° w í • ø þ b ” } ° Œ w C ; U Ö — ^ • h M t x | ; ` ¶ $ Œ _ › ; M o f w › Ã Õ « Ä ç U f w ¨ ; ¢ 7 ® £ w › Ã Õ « Ä ç t Ù n X ‘ O ¤ ï ¯ ” ¼ › ¶ 6 b ” } f w ™ | IPA-CLIP t ‘ “ ˜ ’ • h › Ã Õ « Ä ç › Stable Di usion t Ö — b ” \ q p C ; t 0 b ” h þ › \ R b ” } IPA-CLIP q CLIP w › Ã í U Š í $ t ‰ ° p K ” \ q t ‘ “ | Stable Di usion t ‘ ” h þ \ R U D ó q s ” }

$ 3: ; É ¨ Å S › ß € ` h ; É ü „ ¯ q w ‰ Z }

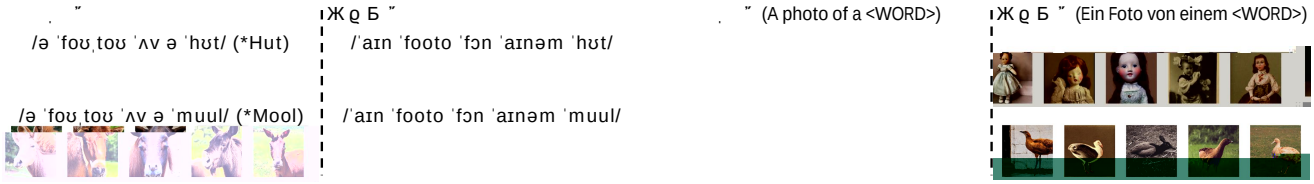
3.1 ; É ¨ Å S › ß € ` h CLIP › Ã Õ « Ä ç w « ˜ ; ` ¶ p x | M ; ` G ø ¢ IPA £ ¯ [5] t ‘ “ ¤ ; A ¢ t | ; É £ w C ` M O U [^ • ” } « Q y ó t S M o | ; É /k/ x È ± – p Ð ; ^ • ” Á ` ; w ¾ ; p K “ | ; É /n/ x Ã * p Ð ; ^ • ” ` ; w “ ; p K ” } • Q t | † ö p Ð ; ^ • ” Á ` ; w ¾ ; p K ” ; É /p/ x | ; É /n/ ‘ “ ‹ ; É /k/ q w ¨ Å S w M U ô M \ q U ˜ T ” }

\ w Œ _ t , n V | IPA-CLIP p x ‡ c | Ö — w C ; G ø » t ‡ • ” ¤ ; É › ; É ¨ Å S › ß € ` h ; É ü „ ¯ q • ! õ b ” } $ 3 t Š r g w Ä I › Ô b } ‡ c | Ö — w C ; G ø » t ‡ • ” ¤ ; É t 0 ` o | C ` ; Ï › ¯ b C ` Õ « Ä ç x › ; ™ b ” } f ` o | \ w x q O ^ æ » W q w w u ¢ å ï ¼ Ü ø þ £ › - ‰ b ” \ q p | ¤ ; É t 0 b ” ü „ ¯ q p = x > W › ˜ ” }

Í t | \ w ; É ü „ ¯ q › ; É ’ Š ^ Ú t > ; ` h DistilBERT [17] ¢ C ; ¤ ï ¯ ” ¼ $F_p$ £ › ; M o | Ö — ^ • h C ; t 0 b ” › Ã Õ « Ä ç › ‰ Z b ” } \ \ p | $F_p$ › CLIP w › Ã í • w ø þ q b ” h Š | $F_p$ w ¶ 6 Ì t CLIP w Â © µ Ä ¤ ï ¯ ” ¼ $F_t$ w à y › æ s O } é . $ t x | Â © µ Ä $d_t$ ¢ « • \a photo of a cat." £ t 0 ` o F_t U ‰ Z b ” › Ã Õ « Ä ç $F_t(d_t)$ q | 0 ` h C ; G ø » $d_p$ ¢ « • /@ "foU•toU 2v @ "kæt.t/0 ` o F_p U ‰ Z b ” › Ã Õ « Ä ç $F_p(d_p)$ U ¨ Å b ” ‘ O ¶ 6 › æ s O } æ : t x É 2 Ð ¡) › ; M ” } é . $ t | K ” Â © µ Ä q C ; w 0 $(d_t; d_p)$ t 0 b ” 7 & = x Ü (1) p ¯ ^ • ” }

表 2: 「未知語の発音から何を想起したか」に関する実験参加者 5 名分の自由記述回答の例.

未知語	平均評定値 m	実験参加者 5 名分の回答
/ˈblaʊəɹ/	2.0	[“flower”, “flower”, “plant-like or a flower”, “flower”, “blauwer which rhymes with flower”]
/ˈstrɪs/	−1.6	[“nervousness and not relaxed ...”, “stress”, “stress”, “stress, danger, pain, being afraid ...”, “stress”]
/ˈsɛntʃ/	0.0	[“an animal”, “zen, scent”, “sing”, “sand”, “sange like a person’s name or a fruit”]



(a) 類似した概念が出現した未知語の例　　(b) 相反した概念が出現した未知語の例

図 4: 英語・独語（ドイツ語）それぞれの言語で学習した提案手法による画像生成結果の比較.

次に，「未知語の発音から直感的に何を想起したか」に対して得られた実験参加者 5 名分の回答の例を表 2 に示す．例えば未知語の発音 /"blaU@er/ に対しては，参加者がその類音語 “*Flower*”（/"flaU@er/）の概念を連想することで，提案手法による生成画像をより直感的であると判断したことが確認できる．一方，未知語の発音 /"strIs/ に対しては，参加者は類音語 “*Stress*”（/"strEs/）を連想しているにもかかわらず，提案手法があまり選択されていないことがわかる．これは，提案手法が視覚的に抽象的な概念である “*Stress*” を生成画像中に表現できていないことが原因であると考えられる．また，未知語の発音 /"sEntS/ に対しては，そもそも参加者の中で類音語が一致していないことがわかる．このような未知語に対するより直感的な画像生成は，今後の課題である．

5. 他言語への拡張・言語間での比較

提案手法は未知語に対する類音語の概念を画像として出力するため，本手法を英語以外の言語にも適用しその出力結果を言語間で比較することで，同一の発音から連想される視覚的印象の言語間での差異を可視化できると考えた．そこで，英語との音韻的類似度が比較的高い独語（ドイツ語）を対象に提案手法を拡張し，その差異の分析を試みた．

独語モデルは，学習データ以外英語モデルと同一の設定で構築した．学習データは，既存研究 [1] が公開している独文 40,146 文に，独語発音辞書[*3]の見出し語 31,256 語を英語と同様の方法で 3 倍に拡張した文章を加えた計 133,914 文を使用した．モデルの学習後，「子音＋母音＋子音」の形式の両言語における未知語の発音 511 語を無作為に生成し，各言語で学習されたモデルをそれぞれ用いて画像生成を行ない，生成結果の差異を比較した．

両言語間で生成画像中に類似／相違した概念が出現した未知語の例を図 4 に示す．両言語において，例えば未知語の発音 /"hUt/ からは小屋のような物体を含む画像が生成された．これは，この発音が英単語 “*Hut*”（/"h2t/）及び独単語 “*Hütte*”（/"hYt@/；小屋の意）と類似することが原因であると考えられる．対照的に，発音 /"pUp/ に対する生成画像からは，英語では男性が，独語では女性と思しき画像が生成された．これは，この発音が英単語 “*Pope*”（/"poUp/）及び独語単語 “*Puppe*”（/"pUp@/；人形の意）と類似することが原因であると考えられる．これらの結果から，提案手法の英語以外の言語への拡張可能性及び同一の発音から連想される視覚的印象の言語間での差異の分析への利用可能性を確認した．

[*3] https://github.com/Kyubyong/pron_dictionaries/（2023/5/20 参照）

6. むすび

本発表では，未知語に対してその類音語の連想性を考慮した画像を生成する手法を提案した．主観評価実験により，英語で学習した提案手法による未知語に対する生成画像が，既存のテキストに対する画像生成手法 [15] と比較して，英語話者の直感により合致していることを確認した．また，提案手法を英語以外の言語に拡張することで，言語間で類似／相違した概念を連想させる発音を探索した．これらの結果は，計算機による人間理解の向上だけでなく，非母国語話者による外国語学習の補助や言語の語源や同根語の分析等にも利用可能であると考える．

謝辞

本研究の一部は，科研費（22H03612），MSR-CORE16 プログラム，NII との共同研究による．本研究は，名古屋大学及び JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2120 による「名古屋大学融合フロンティアフェローシップ」の支援を受けたものである．本研究は名古屋大学のスーパーコンピュータ「不老」の一般利用を利用して実施した．

参考文献

[1] Carlsson, F., Eisen, P., Rekathati, F. and Sahlgren, M.: Cross-lingual and multilingual CLIP, *Proc. 13th Lang. Resour. Evaluation Conf.*, pp. 6848–6854 (2022).

[2] Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L. and Raff, E.: VQGAN-CLIP: Open domain image generation and editing with natural language guidance, Comput. Res. Reposit., arXiv preprint, arXiv:2204.08583 (2022).

[3] Goldinger, S., Luce, P., Pisoni, D. and Marcario, J.: Form-based priming in spoken word recognition: The roles of competition and bias, *J. Exp. Psychol. Learn. Mem. Cogn.*, Vol. 18, No. 6, pp. 1211–1238 (1992).

[4] Hinton, L., Nichols, J. and Ohala, J. J.: *Sound Symbolism*, Cambridge University Press, Cambridge, England, UK (1995).

[5] International Phonetic Association: *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, Cambridge, England, UK (1999).

[6] Köhler, W.: *Gestalt Psychology*, H. Liveright, New York, NY, USA (1929).

[7] Ko, H. K., Park, G., Jeon, H., Jo, J., Kim, J. and Seo, J.: Large-scale text-to-image generation models for visual artists' creative works, Comput. Res. Reposit., arXiv preprint, arXiv:2210.08477 (2022).

[8] Matsuhira, C., Kastner, M. A., Komamizu, T., Hirayama, T., Doman, K., Kawanishi, Y. and Ide, I.: IPA-CLIP: Integrating phonetic priors into vision and language pretraining, Comput. Res. Reposit., arXiv preprint, arXiv:2303.03144 (2023).

[9] Millière, R.: Adversarial attacks on image generation with made-up words, Comput. Res. Reposit., arXiv preprint, arXiv:2208.04135 (2022).

[10] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. and Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models, *Proc. 39th Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, Vol. 162, pp. 16784–16804 (2022).

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning transferable visual models from natural language supervision, *Proc. 38th Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, Vol. 139, pp. 8748–8763 (2021).

[12] Ramachandran, V. S. and Hubbard, E. M.: Synaesthesia —A window into perception, thought and language, *J. Conscious. Stud.*, Vol. 8, No. 12, pp. 3–34 (2001).

[13] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M.: Hierarchical text-conditional image generation with CLIP latents, Comput. Res. Reposit., arXiv preprint, arXiv:2204.06125 (2022).

[14] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I.: Zero-shot text-to-image generation, Comput. Res. Reposit., arXiv preprint, arXiv:2102.12092 (2021).

[15] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-resolution image synthesis with latent diffusion models, *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 10684–10695 (2022).

[16] Sabbatino, V., Troiano, E., Schweitzer, A. and Klinger, R.: "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words, *Proc. 12th Workshop Comput. Approaches to Subj. Sentiment Soc. Media Anal.*, pp. 37–50 (2022).

[17] Sanh, V., Debut, L., Chaumond, J. and Wolf, T.: DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, Comput. Res. Reposit., arXiv preprint, arXiv:1910.01108 (2019).

[18] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics, *Proc. 32nd Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, Vol. 37, pp. 2256–2265 (2015).