Society Column

When Creative AI Meets Conversational AI

Xianchao Wu[†]

1 Introduction

Following the rapid developing of artificial intelligence (AI) boosted by deep neural networks, we are wondering how far is it going for AI to be really *creative*. We know that human beings are intelligent in terms of creating novel things from scratch. Art can be selected as one type of *Turing-Test*: given two paintings, can we distinguish which one is painted by AI and which one is by a real-world person? This same question stands for music creating, poem writing, singing and so on. We further ask ourselves, can we be involved in the whole AI creation process? Say, we communicate with AI to exchange ideas of what's the next painting topic or genre, or what's the next movement of the music. Through these interactive AI creation processes, we hope to inspire AI and be inspired by AI, just alike we learn from Alpha-Go[1] after it learnt from us first and then by itself through self-supervised learning. For Go gaming, we know the rules of win or lose. However, for AI creation, evaluation metrics are subjectively defined in creative ways.

In this year's (27th) Natural Language Processing (NLP 2021) conference, we proposed and organized the first workshop named "when creative AI meets conversational AI",[2] or, briefly "CAI+CAI=CAI".[2] In this one-day workshop, there are two technical papers and eight invited talks. These two technical papers respectively and interestingly cover one direction of text-to-image leveraging DM-GAN (Zhu et al. 2019) and ManiGAN (Li et al. 2020) for bird image generation (Azuaje et al. 2021) and its reversed direction of image-to-text by visual-text integrated Transformer (森 他 2021) for story generation. The eight invited talks cover most major directions of creative AI and conversational AI, such as voice conversion (Zhao et al. 2020), image generation (Jiang et al. 2021), contents retrieving (Yu et al. 2019), fine-art paintings (Huckle et al. 2020), image-to-image translation (Guo et al. 2021), AI painting, music generating and singing (Wu et al. 2020), poem-writing (Wu et al. 2017), and multi-modal learning for medical and healthcare applications (Ma et al. 2021; Obinata et al. 2020).

[†] NVIDIA, xianchaow@nvidia.com

[1] `https://deepmind.com/research/case-studies/alphago-the-story-so-far`

[2] `https://sites.google.com/view/cai-workshop`

2 Motivation of This Workshop

Our workshop is inspired by the following facts. Creative AI, training *generative* deep neural networks for NLP (such as poems, Haiku, stories), image (such as painting, animation), and speech (such as classic and popular music generation, singing), has achieved impressive milestones during recent years, thanks to deep neural networks such as attentive encoder-decoder architectures alike Transformers (Vaswani et al. 2017), generative-discriminative frameworks alike GANs (Goodfellow et al. 2014) and self-supervised encoders alike VAEs (Razavi et al. 2019). In industry, conversational AI products such as Apple's Siri, Microsoft's Cortana, Google Home, Amazon's Alexa, XiaoICE (Zhou et al. 2020), Rinna (Wu et al. 2016) that support text and speech based multi-modal communication between chatbots and human beings, have obtained millions of users in Japan and billions of users globally.

In order to construct the strong *persona* of conversational AI products, chatbots are enhanced to be able to interactively write poems, create songs, sing, and even tell stories, through multi-turn communications with end users. Furthermore, QA-style and IR-oriented chatbots of general domains and vertical domains such as finance, healthcare and even emotional pure-chatting are also requiring generative, creative and explainable AI models to support the multi-modal and multi-turn interactions with human beings.

In this workshop, we were aiming at collecting, sharing, and discussing state-of-the-art research on creative AI and conversational AI, empowered by large-scale open datasets, open-source architectures, and distributed GPU platforms. Most importantly, with creative AI combined with conversational AI, we are aiming at bringing AI to help assisting under-represented groups' learning and communicating with the real world, such as interactive music therapy, children's painting guiding, emotional caring for social phobia and elderly cognition guarding. These goals derive our current and future CAI^2 workshops.

3 Roadmap of Technical Directions

Figure 1 illustrates three major directions of creative AI,[3] text, image and speech and further their combinations. First, GANs and generative large-scale pretrained language models such as GPT-3 have been employed for text generating. The up-left corner also shows a Haiku generated by Rinna (Wu et al. 2017). Second, GANs and visual-transformers have been utilized for

[3] I also gave a same title speech in GTC 2021: S31384 `https://gtc21.event.nvidia.com/media/When%20Creative%20AI%20Meets%20Conversational%20AI%20%5BS31384%5D/1_1nr6o531`

controllable image generations. There are eight images for kerchiefs generated by inspiring customers' individual favours in the middle-left corner. Third, music and speech directions are of enormous market scale where AI has been investigated for piano generating that designed time-valued note tuples as inputs to transformer-xl (Dai et al. 2019) with joint pitch/velocity/melody learning (Wu et al. 2020) and singing voice synthesis (e.g., popular music, Peking opera) using non-autoregressive frameworks such as Transformer's encoders following fastspeech (Lu et al. 2020), as shown in the bottom-left corner of Figure 1.

These three directions can be combined together for multi-modal or cross-modal AI creation. First, text and image can be combined for directions such as text-guided painting. In the example given in the up-right corner of Figure 1, the input text is "*urbanization in China*" and there are six paintings respectively drawn by historically famous artists such as Franz Marc, Rembrandt, Edvard Munch, Emile Bernard, van Gogh, and Gauguin. It is interesting to ask artists that

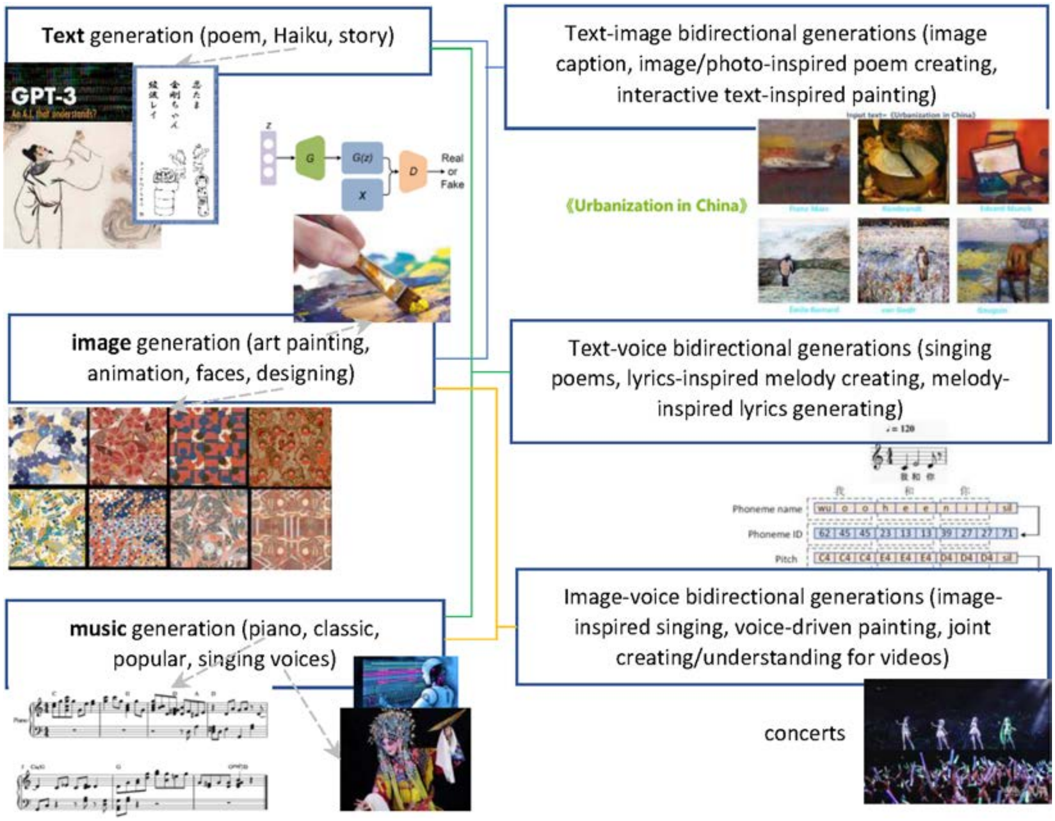
Text generation (poem, Haiku, story)
GPT-3
An A.I. that understands?
z
G
G(z)
x
D
Real or Fake
image generation (art painting, animation, faces, designing)
music generation (piano, classic, popular, singing voices)
Text-image bidirectional generations (image caption, image/photo-inspired poem creating, interactive text-inspired painting)
Input text= 《Urbanization in China》
《Urbanization in China》
Text-voice bidirectional generations (singing poems, lyrics-inspired melody creating, melody-inspired lyrics generating)
♩= 120
我 和 你
Phoneme name
Phoneme ID
Pitch
wu o o h e e n i i i sil
62 45 45 23 13 13 39 27 27 71
C4 C4 C4 E4 E4 E4 D4 D4 D4 sil
Image-voice bidirectional generations (image-inspired singing, voice-driven painting, joint creating/understanding for videos)
concerts

Fig. 1 Illustration of creative AI's single/combined directions.

lived centuries before to draw pictures of modern topics. In addition, the painting process is designed to be interactive: users are encouraged to keep guiding every time a painting is drawn by AI so that the images can be kept updating in multi-turns. This is helpful for users to better understand AI's painting process and for beginners to practice following hints from AI. Second, text and voice can be combined so that we can train an AI singer who can sing by given lyrics and melody and chord progress are generated on the fly (Zhu et al. 2018). The example given in right-middle corner is to ask AI to generate singing for a lyric sentence "*I and you*" (Lu et al. 2020). Third, by combining images and voices following timeline, concerts have been hosted by AI singers or virtual singers, considering that there are a list of famous virtual singers.

Besides conversational AI products developed by a list of AI companies, there are open-source developer-oriented conversational AI platforms, such as NVIDIA's Jarvis[4] which includes 35 pretrained models of speech recognition, NLP (NER, QA, user intention classification, slot filling), and speech synthesis functions. Jarvis is a fully accelerated application framework for building multi-modal conversational AI services that use an end-to-end deep learning pipeline. Developers at enterprises can easily fine-tune state-of-the-art models on their data to achieve a deeper understanding of their specific context and optimize for inference to offer end-to-end real-time services that run in less than 300 milliseconds and delivers 7x higher throughput on GPUs compared with CPUs. Combining creative AI with these open-source conversational AI platforms can possibly save researchers and developers time for idea and product implementing.

4 Conclusion

Human being constructed modern industrialized world since we are creative and converse with each other for cooperation. Now, we are empowering AI systems to be more and more creative so that we can cooperate with them as well through numerous multi-modal channels. Assisted by AI algorithms, big data and cloud computing (ABC), CAI^2 is absorbing more focus and is supposed to contribute continually and impressively to a better society.

Acknowledgement

The author would like to acknowledge the large contributions of Dr. Gang Niu and Dr. Lin Gu from RIKEN, Assistant Professor Haitao Yu from University of Tsukuba, Dr. Yi Zhao (formerlly)

[4] https://developer.nvidia.com/nvidia-jarvis

from NII, Dr. Yulan Yan from IBM and my colleagues Dr. Peiying Ruan, Mana Murakami and Khanh Vo Duc from NVIDIA. We also express our thankfulness to the technical paper authors and the invited speakers.

References

Azuaje, G., Liew, K., Yada, S., Wakamiya, S., Aramaki, E., and Khan, D. (2021). "Birdscribe: A Semantic Writing Assistant Employing Text-based Image Generation and Modification." In *1st Workshop on When Creative AI Meets Conversational AI*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context." In *Proceedings of ACL*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). "Generative Adversarial Nets." In *NeurIPS*, Vol. 27.

Guo, J., Li, J., Gong, M., Fu, H., Zhang, K., and Tao, D. (2021). "Minimal Geometry-Distortion Constraint for Unsupervised Image-to-Image Translation." In *OpenReview*.

Huckle, N., Garcia, N., and Nakashima, Y. (2020). "Demographic Influences on Contemporary Art with Unsupervised Style Embeddings." In *Computer Vision - ECCV 2020 Workshops*, Vol. 12536 of *Lecture Notes in Computer Science*, pp. 126–142. Springer.

Jiang, Y., Chang, S., and Wang, Z. (2021). "TransGAN: Two Transformers Can Make One Strong GAN." *CoRR*, **abs/2102.07074**.

Li, B., Qi, X., Lukasiewicz, T., and Torr, P. H. S. (2020). "ManiGAN: Text-Guided Image Manipulation." In *Proceedings of CVPR 2020*. IEEE Computer Society.

Lu, P., Wu, J., Luan, J., Tan, X., and Zhou, L. (2020). "XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System.".

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., and Lu, F. (2021). "Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems." *Pattern Recognition*, **110**, p. 107332.

森友亮，上原康平，原田達也 (2021). 視覚・言語融合 Transformer モデルによる画像からの物語文生成. In *1st Workshop on When Creative AI Meets Conversational AI*. [Y. Mori et al. (2021). VisualNT-BART: Image to Narrative Generation with Vision and Language Transformer. 1st Workshop on When Creative AI Meets Conversational AI.].

Obinata, H., Ruan, P., Mori, H., Zhu, W., Sasaki, H., Tatsuya, K., Wakana, M., Tanaka, M., Hsu, P.-L., Yang, D., Xu, Z., Xu, D., Tamura, K., and Yokobori, S. (2020). "Can Artificial Intelligence Predict the Need for Oxygen Therapy in Early Stage COVID-19 Pneumonia?"

In *In researchsquare.com.*

Razavi, A., van den Oord, A., and Vinyals, O. (2019). "Generating Diverse High-Fidelity Images with VQ-VAE-2." *CoRR*, **abs/1906.00446**.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is All you Need." In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

Wu, X., Ito, K., Iida, K., Tsuboi, K., and Klyen, M. (2016). りんな：女子高生人工知能. 言語処理学会第 22 回年次大会発表論文集, pp. 306–309. [X. Wu et al. (2016). Rinna: High School Girl Artificial Intelligence. Proceedings of the 22nd Annual Meeting for the Association for Natural Language Processing, pp. 306–309.].

Wu, X., Klyen, M., Ito, K., and Chen, Z. (2017). "Haiku Generation Using Deep Neural Networks." In *Proceedings of the 23th Natural Language Processing.*

Wu, X., Wang, C., and Lei, Q. (2020). "Transformer-XL Based Music Generation with Multiple Sequences of Time-valued Notes." *CoRR*, **abs/2007.07244**.

Yu, H., Jatowt, A., Joho, H., Jose, J. M., Yang, X., and Chen, L. (2019). "WassRank: Listwise Document Ranking Using Optimal Transport Theory." In Culpepper, J. S., Moffat, A., Bennett, P. N., and Lerman, K. (Eds.), *Proceedings of WSDM 2019*, pp. 24–32. ACM.

Zhao, Y., Li, H., Lai, C.-I., Williams, J., Cooper, E., and Yamagishi, J. (2020). "Improved Prosody from Learned F0 Codebook Representations for VQ-VAE Speech Waveform Reconstruction." *CoRR*, **abs/2005.07884**.

Zhou, L., Gao, J., Li, D., and Shum, H.-Y. (2020). "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot." *Computational Linguistics*, **46** (1), pp. 53–93.

Zhu, H., Liu, Q., Yuan, N. J., Qin, C., Li, J., Zhang, K., Zhou, G., Wei, F., Xu, Y., and Chen, E. (2018). "XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music." In *Proceedings of SIGKDD*, KDD '18, pp. 2837–2846.

Zhu, M., Pan, P., Chen, W., and Yang, Y. (2019). "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis." In *Proceedings of CVPR*, pp. 5802–5810.

Xianchao Wu: received his Ph.D. degree in Computer Science from The University of Tokyo in 2010, and a master's and bachelor's in Computer Science from Jilin University in 2004 and 2001, respectively. He also has an MBA degree of finance from Hitotsubashi University in 2020. He is currently a senior solution architect and data scientist at NVIDIA. He worked at Baidu until 2015 and

then Microsoft until 2020. His research interests include large-scale pretrained language models, conversational AI, creative AI and financial NLP. He is an inventor and co-inventor of more than 90 patents.