11-877 Advanced Multimodal Machine Learning **Spring 2022**

Week 14: Multimodal Generation and Ethical Concerns

Instructors: L.-P. Morency, A. Zadeh, P. Liang *Synopsis Leads: Justin Lovelace, Dong Won Lee*

Edited by Paul Liang *Scribes: Chonghan Chen, Yuanxin Wang*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources.

In week 14's discussion session, the class aimed to address the challenges in multimodal generation. Topics of interests included the technical challenges in multimodal generation, the evaluation of generation quality, and potential ethical issues of generative models. The following was a list of provided research probes:

1. What are some challenges in multimodal generation beyond generating each modality individually? How can we synchronize generation across multiple modalities?
2. What degree of multimodal modeling is required for these cross-modal generation to be possible? For example, how much do models need to learn regarding cross-modal interactions, alignment, reasoning, etc?
3. What are the qualities we should consider when evaluating outputs from multimodal generation? What do you think is the best practice to evaluate these qualities? Can we efficiently evaluate these qualities, at scale?
4. What are the opportunities and challenges of automatic and human evaluation? How can we combine the best of both worlds?
5. What are the real-world ethical issues regarding generation? How are these risks potentially amplified or reduced when the dataset is multimodal, with heterogeneous modalities? Are there any ethical issues that are specific to multimodal generation?
6. How can we build a taxonomy of the main ethical concerns related to multimodal generation?
7. How can we update our best practices to help address these ethical concerns? Who is better placed to start this dialogue? How can we make significant changes in this direction of reducing ethical issues?

As background, students read the following papers:

1. (Required) VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning [Chen et al., 2021]
2. (Required) Zero-Shot Text-to-Image Generation (DALL-E) [Ramesh et al., 2021]
3. (Required) Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL-E 2) [Ramesh et al., 2022]
4. (Required) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? [Bender et al., 2021]
5. (Suggested) The social impact of deepfakes [Hancock and Bailenson, 2021]
6. (Suggested) What a machine learning tool that turns Obama white can (and can't) tell us about AI bias [Vincent, 2020]
7. (Suggested) What comprises a good talking-head video generation?: A survey and benchmark [Chen et al., 2020]
8. (Suggested) Defending against neural fake news [Zellers et al., 2019]
9. (Suggested) Lessons from the PULSE Model and Discussion [Kurenkov, 2020]

10. (Suggested) Text-to-Image Generation Grounded by Fine-Grained User Attention [Koh et al., 2021]
11. (Suggested) Training for Diversity in Image Paragraph Captioning [Melas-Kyriazi et al., 2018]
12. (Suggested) Multimodal Abstractive Summarization for How2 Videos [Palaskar et al., 2019]
13. (Suggested) Extracting Training Data from Large Language Models [Carlini et al., 2021]
14. (Suggested) What Makes Tom Hanks Look Like Tom Hanks [Suwajanakorn et al., 2015]
15. (Suggested) Video Generation From Text [Li et al., 2018]

We summarize several main takeaway messages from group discussions below:

1 Multimodal Generation Tasks

Table 1: Different tasks in multimodal generation.

Task	Description
Summarization	Summarizing the information from some multi-modal input. For instance, generating a concise textual summary of an audio-visual news broadcast.
Translation	Translating one modality to another. Generating images based on textual descriptions is a widely studied example.
Creation	This involves going from some small input specification to a larger, multimodal output. For example, generating an audio-visual video based on a short textual description. Such tasks receive less attention, likely because they are more challenging.

We outline three broad categories of tasks that fall under the umbrella of multimodal generation: summarization, translation, and creation. We provide short descriptions of these tasks in Table 1.

Summarization involves summarizing the information from some large, multimodal input in a more concise, easily digestible form. For example, providing a concise textual summary of an audio-visual news broadcast would be one example of multimodal summarization. The summary itself could also be multimodal by incorporating relevant visual elements from the broadcast into the summary.

Translation is perhaps the most commonly studied of the three categories. It involves generating one modality from another. A common example is generating images from textual descriptions like DALL·E 2 [Ramesh et al., 2022].

Creation typically involves going from some small, concise input to a larger, more complex multimodal output. An example would be generating some audio-visual video based on a textual description. This area has received the least attention likely because it is the most challenging of the three discussed. For instance, ensuring the coherence of audio and visual indicators for spoken dialogue would be very difficult.

Across the different generation tasks, going beyond textual descriptions and coordinating between multiple modalities for generation is an important area for further study. For instance, a generation method could take an audio description as input or a language-visual input.

2 Challenges in Multimodal Generation

We also outline several core challenges in multimodal generation: controllability, compositionality, synchronization, and capturing long tail phenomena. We provide short descriptions of these in Table 2.

Controllability is desirable in multimodal generation models since a user may want to generate photo-realistic images in a specific artistic style. For controllability, potential approaches include explicitly guiding the decoding process through latent variable models or exploring different sampling schemes.

Multimodal generation models should also be able to understand complex, compositional inputs and appropriately generate compositional outputs across different modalities. The recent Winoground challenge demonstrates that many language-visual models fail to correctly interpret non-standard compositional language and images [Thrush et al., 2022]. Strong unimodal language models have made significant progress at

Table 2: Challenges in multimodal generation.

Challenges	Description
Controllability	Fine-grained control over generation models is desirable. For instance, a user may wish to generate an image in a certain style.
Compositionality	Multimodal generation models need to handle complex, compositional inputs. Recent work has shown that current multimodal methods often fail to handle non-standard compositions (e.g. distinguishing "a lightbulb surrounding some plants" from "some plants surrounding a lightbulb ").
Synchronization	For creation tasks that involve generating multiple modalities, the modalities must be synchronized. For example, the audio of people speaking in a video must match the movement of their mouths.
Capturing long tail phenomena	Many generative models suffer from some form a mode collapse and fail to generate rare, but valid, phenomena. A generative model would ideally be able to generate unique compositions that are not well-represented in their training set.

understanding compositional language [Brown et al., 2020], but a wider gap seems to appear when shifting to the multimodal setting. Figuring out how to close this gap presents an interesting research challenge.

Synchronization is another important concern for creating tasks where multiple modalities are being generated. It is critical that modalities are synchronized appropriately for generation to be semantically coherent. For instance, generating dialogue in a video involves synchronizing a unique voice with a person speaking in the clip. The voices would also need to remain consistent across utterances. Evaluation of synchronization is an important challenge because developing automated evaluation is non-trivial. Human evaluation is a possible alternative, but that comes with its own drawbacks such as a lack of reproducibility.

Effective multimodal generation models should be able to capture long tail phenomena that may not be well-represented within their training dataset. This challenge is related to the problem of mode collapse that has been observed with generative models such as GANs [Goodfellow et al., 2014a]. To handle the last challenge, it is important to generate diverse images, while still remaining faithful to the text that is prompting the generation. Language is very expressive and presents a natural way to represent different modalities. Language descriptions can be left intentionally vague or can be very specific depending on the intent of the user. Utilizing complex, highly specific language descriptions for generation will, however, also increase the difficulty of the generation task. Exploring whether alternative representations (e.g. symbolic) could be developed to guide generative models with greater control is an interesting research direction.

3 Representations for Multimodal Generation

3.1 Self-supervised representations

Many previous works map the input modalities into a single joint latent space, which is later used to generate the desired output modality. Recently, methods such as Data2Vec [Baevski et al., 2022], and Merlot Reserve [Zellers et al., 2022], rely on self-supervised multimodal learning, where the core idea is the predict a masked portion of the joint latent representations of the full input data, similar to that of a masked language modeling task in NLP. Though these methods show tremendous success, they require extremely large training datasets (20 million videos, 1 billion frames) and significant compute resources for training.

3.2 Coordinated representations

The human brain seems to have multiple representations working simultaneously, and it is highly unlikely that our brain's representation of multiple modalities is simply a d-dimensional vector. (1) Is there a different way to represent each modality and (2) coordinate them in a way like our brain does? How do we generate coherent output with coordinated representations (e.g. brain generates 3D scene from vision). Is it possible to have joint generators instead of separate generators, so that we can better synchronize generation across modalities?

3.3 Structured representations

What kind of structure is appropriate for each generation task (i.e., image generation, video generation)? Rather than feeding in raw images as input, investigating structured input representations such as knowledge graphs or scene graphs, could be promising directions of research [Johnson et al., 2018].

4 Evaluating Generated Content

A key challenge in generation problems is evaluation. In many specific applications, including but not limited to speech synthesis [Cambre et al., 2020], gesture generation [Wolfert et al., 2022], and language modelling [Papineni et al., 2002, Lin, 2004, Banerjee and Lavie, 2005, Vedantam et al., 2015], codifying an all-encompassing evaluation metric that can capture naturalness and humanlikeness has been a core research direction. Separately, (1) designing a differentiable loss term, (2) an evaluation metric for hyper parameter tuning, and (3) human studies to measure whether the generation models are exhibiting desired properties are all important research problems. Furthermore, increasing efforts are being made to reduce the gap between automatic and subjective evaluations. For instance, in speech representation learning, HuBERT [Hsu et al., 2021], offers 3 different metrics of target quality (phone purity, cluster purity, phone-normalized mutual information).

In earlier works, carefully designing losses or reward functions that penalize/reward certain behaviors were a core line of research. A way to alleviate defining specific desired properties is to learn this function directly from data, which is specifically what a GAN [Goodfellow et al., 2014b] does. Now the question then lends itself to, what does a discriminator really learn? Could we extract the properties that define a good generator by carefully examining the discriminator? An approach to do this would be to reverse engineer a neural network as proposed by a group of researchers at Open AI [Cammarata et al., 2020].

5 Biases in Generation

Before actually deploying generation models, we need to carefully consider the ethical considerations surrounding their use. Language is a social construct, as a result, it will inevitably contain biases. Consequently, in large-scale language models, such as GPT-3, we see negative associations with race, gender, and religion. Furthermore, GPT systems have been shown to encode harmful bias across identities, which include abusive language [Bender et al., 2021, Abid et al., 2021, Brown et al., 2020]. The problem of encoding biases regarding demographic groups has also been found in multimodal generation [Mishkin et al., 2022]. Furthermore, existing measures intended to prevent undesireable behavior can often be circumvented. For example, although DALL·E 2 prohibits the use of the word "blood", the phrase "a pool of red liquid" can be used to generate an image that looks like it has a pool of blood [Mishkin et al., 2022].

Recently, OpenAI released PALMS [Solaiman and Dennison, 2021], which describes a process to improve model behavior by crafting and fine-tuning on a dataset that reflects a predetermined set of target values. Such methods that can be used to quickly adapt models to reflect societal values is important. Moreover, developing models that are aware of biases in their predictions should be prioritized as the next step.

References

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. *arXiv preprint arXiv:2102.10407*, 2021.

Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014a. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014b.

Jeffrey T Hancock and Jeremy N Bailenson. The social impact of deepfakes, 2021.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 237–246, 2021.

Andrey Kurenkov. Lessons from the pulse model and discussion. *The Gradient*, page 11, 2020.

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Luke Melas-Kyriazi, Alexander M Rush, and George Han. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 757–761, 2018.

Pamela Mishkin, Lama Ahmad, Miles Brundage, Gretchen Krueger, and Girish Sastry. Dall·e 2 preview - risks and limitations. 2022. URL https://github.com/openai/dalle-2-preview/blob/main/system-card.md.

Shruti Palaskar, Jindřich Libovickỳ, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34, 2021.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. What makes tom hanks look like tom hanks. In *Proceedings of the IEEE international conference on computer vision*, pages 3952–3960, 2015.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL https://arxiv.org/abs/2204.03162.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

James Vincent. What a machine learning tool that turns obama white can (and can't) tell us about ai bias. *Retrieved October*, 28:2021, 2020.

Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, 2022.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. *arXiv preprint arXiv:2201.02639*, 2022.