11-877 Advanced Multimodal Machine Learning **Spring 2023**

Week 10: Multimodal Generation and Ethics

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Gaoussou Kebe, Suzanne Nie*

Edited by Paul Liang *Scribes: Leena Mathur, Yiqing Xie*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

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 10's discussion, the class discussed challenges that arise with multimodal generation, evaluation, and ethics. The following was a list of provided research probes:

1. 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?
2. What are some challenges in multimodal generation beyond generating each modality individually? How can we synchronize generation across multiple modalities?
3. What aspects of multimodal are prerequisites for generation to be possible? For example, how much do models need to learn regarding heterogeneity, connections, and interactions?
4. There have been many directions towards conditional generation without fully paired data, or paired data at more coarse granularities (e.g., text-video generation using only text-image data). What is a taxonomy of weak supervision approaches for generation? How do we know what type of data is necessary for accurate generation?
5. What are the opportunities and challenges of automatic and human evaluation? How can we combine the best of both worlds?
6. 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?
7. How can we build a taxonomy of the main ethical concerns related to multimodal generation?
8. 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) Show me what and tell me how: Video synthesis via multimodal conditioning [Han et al., 2022].
2. (Required) Grounding language models to images for multimodal generation [Koh et al., 2023].
3. (Suggested) Visual ChatGPT: Talking, drawing and editing with visual foundation models [Wu et al., 2023].
4. (Suggested) Make-A-Video: Text-to-Video Generation without Text-Video Data [Singer et al., 2022].
5. (Suggested) Holistic Evaluation of Language Models [Liang et al., 2022].
6. (Suggested) What comprises a good talking-head video generation?: A Survey and Benchmark [Chen et al., 2020].
7. (Relevant) A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity [Bang et al., 2023].
8. (Relevant) It's Raw! Audio Generation with State-Space Models [Goel et al., 2022].

9. (Relevant) Equivariant Diffusion for Molecule Generation in 3D [Hoogeboom et al., 2022].

We summarize several main takeaway messages from group discussions below:

1 Evaluation (Probes 1, 5)

In Table 1 we outline different factors to consider when evaluating language generation.

Type	Definition
Relevance	How sensitive are outputs to semantic perturbations of the input?
Consistency	How well do outputs match inputs?
Diversity	How diverse are outputs?
Robustness	How sensitive is the model to noisy (non-semantic) perturbations?
Factuality	Does the model give true information?
Realisticness	How similar are model outputs to human language?
Bias	Does the model carry over bias from its training data?
Toxicity	Is there a guard against generating toxic outputs?

Table 1: Factors to consider when evaluating language generation.

Automatic evaluation frameworks have several benefits including efficiency, reproducibility, ease to quantitatively compare model performance, and objectivity. We discussed a few specific methods, datasets, and metrics we can use to automatically evaluate generated data.

- For relevance, a dataset of multiple choice question format can be made to select outputs that are most relevant to the inputs.
- For consistency, we can use cyclic reconstruction. For example, for QA tasks, reconstruct the question from the generated answer and then generate another answer to match with the original answer. For text to image models, generate an image, use a state-of-the-art image captioning model, and match the input text to the caption.
- For diversity, we can use entropy of generated outputs as a metric. We can also count how many ground truth outputs were generated; if this is a large proportion, it would show lack of diversity.
- **Bleu score** (bilingual evaluation understudy) [Papineni et al., 2002] The bleu score, between 0 and 1, measures how close a machine translated output matches a professional human translated output.
- **Perplexity** The PPL score measures the most likely next utterance in a dialogue based on the previous conversation turns.
- To detect toxicity, train a discriminator with reported data to filter out toxic outputs.

However, automatic evaluation is more difficult to apply to subjective modalities, such as music generation. Human evaluation frameworks must be used in these scenarios.

2 Challenges in Multimodal Generation (Probe 2)

How can a model choose which modality to generate? For humans, when we text, we naturally switch between text, image, and video modalities to best convey our thoughts. How can we train a model to behave this way? For example, in Koh et al. [2023], in dialogue with the model, it is explicitly asked for an image, but there are situations where the model must select which modality is best to generate without prompt. In a test dataset, references may contain different modalities (i.e. the model's output is text and the reference includes text and video.)

How would we evaluate a model's output given these differences? One method discussed was learning a special token for when one modality ends that also contains information about what the next modality is. Another method discussed was using two special tokens to denote the starts and ends of different modalities. The modality we want the model to output is the one that conveys the most information. Based on a reward function, our goal would be to learn succinctness to choose the modality that requires the least amount of

bits to represent the information. Two challenges we face is how to measure the succinctness of different modalities and how to tackle long sequence generation while still being faithful.

3 Ethical Issues (Probes 6, 7, 8)

What are the real-world ethical concerns regarding multimodal generation? Multimodal generation raises a wide array of real-world ethical issues. In Table 2, we present some of the main ethical issues to consider.

Issues	Definition
Job displacement	How will generative models affect human work and employment?
Plagiarism	How original are the outputs of generative models?
Privacy	How will large generative models use the personal data contained in their training sets?
Human creativity	How will generative models influence human innovation and expression?
Deepfakes	How realistic and harmful are the images, audio, and video generated by models?
Cyberbullying	How will generative models affect the abuse and harassment of users and subjects?
Fake news	How will generative models affect the creation and dissemination of false or misleading information?

Table 2: Ethical issues in multimodal generation.

How can we address these ethical concerns? We discussed these ethical concerns and put forward potential solutions to address them.

- *Job displacement:* Generative models are taking over tasks previously done by humans, which can lead to job displacement [Eloundou et al., 2023]. However, they can also help humans by automating mundane tasks, allowing them to focus on more creative and interesting work. So, we must strive for a responsible and ethical approach to the integration of generative models into the workforce, with a focus on achieving a sustainable balance between automation and human involvement. The legal and policy aspects of AI have not kept pace with the rapid progress of research. To avoid serious job loss issues, policies need to be revised to address these challenges.
- *Plagiarism and Privacy:* Even though generative models can generate creative content, it is still not entirely clear if they can create genuinely original ideas without copying and combining existing human sources. There is currently no foolproof method of identifying the sources used in the generation process, raising tons of copyright issues [Franceschelli and Musolesi, 2022]. We propose that attribution models [Yu et al., 2019] can be used to identify the sources used by models to generate content and digital signatures for authors can be developed to protect their intellectual property. Generative models might also be trained on personal data from the internet or elsewhere, raising privacy concerns. Therefore, there is a need for models that can detect personal information in the training data of a generative model [Hayes et al., 2017] and strategies to measure and reduce the privacy leakage of generative models.
- *Human creativity:* It can be argued that generative models are not replacements for human creativity, but rather tools that can enhance and complement it [Fenwick and Jurcys, 2023]. Humans still have the ultimate control over the generation and evaluation of creative content. However, to address ethical concerns about generative models and human creativity, more research needs to be done on the impact of these models and the role of humans in the creative process.
- *Deepfakes, Cyberbullying, and Fake news:* The realism and potential harm of the images, audio, and video generated by models is another ethical issue. Adversarial models such as Generative Adversarial Networks (GANs) can be used to differentiate between machine-generated and human-generated content. However, the effectiveness of these models depends on the quality of the discriminator used. Another concern with deepfakes is the generation of nudity or other fake sensitive content. Before releasing a powerful generative model to the public, it is necessary to prevent it from being able to generate such information in pretraining, and develop models that can distinguish real multimodal recordings from deepfakes [Groh et al., 2022].

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

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

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Mark Fenwick and Paulius Jurcys. Originality and the future of copyright in an age of generative ai. *Available at SSRN 4354449*, 2023.

Giorgio Franceschelli and Mirco Musolesi. Copyright in generative deep learning. *Data & Policy*, 4:e17, 2022.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Re. It's raw! Audio generation with state-space models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7616–7633. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/goel22a.html.

Matthew Groh, Aruna Sankaranarayanan, Andrew Lippman, and Rosalind Picard. Human detection of political deepfakes across transcripts, audio, and video. *arXiv preprint arXiv:2202.12883*, 2022.

Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022.

Jamie Hayes, Luca Melis, George Danezis, and ED Cristofaro. Logan: Evaluating information leakage of generative models using generative adversarial networks. *arXiv preprint arXiv:1705.07663*, 18, 2017.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8867–8887. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2022.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. 2022.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019.