BIODS 276 / CS 286 Advanced Topics in Computer Vision and Biomedicine
Fall 2024

Assignment 2 - Extending Generative Models

In this assignment, you will write a 1500-2000 word analysis paper, using the NeurIPS LaTeX template. The goal of the assignment is to synthesize and think critically about the topics covered in class, through reading and analyzing recent papers. The assignment must be completed individually; you may not work in groups.

Please select one of the following analysis categories: “Extending generative models to more complex modality setups” or “Diving into generative models in a biomedical application”, and choose two research papers in that category based on the provided guidelines. This course has two analysis assignments, one focused on representation learning models and the other on generative models. Both assignments will have the same analysis categories, and ***whichever category you chose for the first assignment, you must choose the other for this second assignment.***

Option 1. Extending generative models to more complex modality setups:

In lecture, we have focused on vision-only and vision-language *generative models* in a “standard” setting: 2D images are in a standard resolution (less than 1000x1000 pixels), and the text is a short caption string (less than 50 words). So now, dive deeply into generative modeling outside the standard setting, such as non-standard 2d images or vision paired with other (non-language) modalities. Some examples include video, 3D vision-only, high-resolution 2d image-only, joint image-audio, joint video-language, etc.

Choose one such setting and identify two research papers that propose approaches for your selected modality setup. The key requirement is that your chosen modality setup should not be in the ‘standard setting’, but should still involve visual pixel data.

Option 2. Diving into generative models in a biomedical application:

We have surveyed examples of generative models in different biomedical contexts. Now, dive more deeply into generative models for a specific biomedical application or use case of your choice. Identify two research papers that propose generative modeling approaches for your chosen application, using large vision-only or multimodal (vision-X) models. These papers may be ones discussed in class or others you discover independently. You are also not limited to 2D image or image-text models. For example, you may select more complex modality setups such as those described in ‘option 1’ prompt.

Prompt

Given the two selected papers in whichever category you have chosen for this assignment, compare and contrast their approaches, answering these key questions:

1. Briefly give an introduction to the context of your setting and its motivation. Explain why you chose your 2 papers.

2. What are the most significant tasks? Highlight the tasks that are explicitly discussed in the papers, but try to also consider other relevant tasks.
3. Describe the data and what is distinct about it in this context:
    a. For option 1, explain how the data modality of choice has been represented in deep learning, including in your chosen papers as well as in prior work (if applicable). For example, images are represented as pixel values in CNNs, and patches of pixels in vision transformers.
    b. For option 2, explain what is challenging about the data content compared to natural images. For example in lung tumor classification in X-ray, the classification of healthy / disease can depend on a very small part of the image, while for cat/dog classification, that is not the case.
4. For each paper, how do they motivate their methodology? Put another way, what is their key insight?
5. Describe the strengths and weaknesses of each paper.
6. If one paper is stronger than another, why? If it is not clear that one paper is stronger than the other, why?
7. Summarize and synthesize the key conclusions of your analysis.
8. What are important or promising directions for future work?

You may reference additional papers if they help support an argument, but the two chosen papers should have primary focus for the comparative analysis. Your reports will be assessed based on whether you address this list of questions.

Guidelines:
The analysis paper should be written in the format of standard machine learning papers, abridged as stated following to adapt to the assignment goals. The paper should contain the following sections:

- **Prompt Choice, Selected Two Papers (write out the full citation)**
- **(10%) Introduction (~0.5 page).** Prompt question 1.
- **(20%) Problem background (~0.5-1page)** Prompt questions 2-3.
- **(50%) Analysis (2-3 pages).** Prompt questions 4-6. This should be by far the most significant section of your paper. Feel free to break it down into further subsections if you like. Please reference the prompt questions for further details.
- **(20%) Conclusion and future work (0.5 pages).** Prompt questions 7-8.

The report should be 1500-2000 words and in the Neurips latex format (see here and here), not including references. You are encouraged to include figures.

Choosing two papers
Below are suggested papers, but you are welcome to select your own. If you choose your own paper, please make a private post on Ed and tell us the papers so that we can check that it's in scope. We'll get back to you within 48hrs, please plan accordingly with respect to the assignment deadline.

You can use papers that were used in class if they were only reviewed briefly (4 total slides or less).

You should focus on methods for generative modeling. Some papers may discuss some representation learning pre-training phase. That's fine, but your analysis should emphasize the generative modeling parts.

Suggested topics & papers for "Extending to more complex modality setups":

- From the lectures on "Vision Diffusion and Generative Models"
    - Text to video
        - Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning
        - CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer
        - Imagen Video: High Definition Video Generation with Diffusion Models
    - Image to 3D
        - VFusion3D: Learning Scalable 3D Generative Models from Video Diffusion Models
        - LRM: Large Reconstruction Model for Single Image to 3D
    - Style-control in text-to-image generation
        - Style Aligned Image Generation via Shared Attention
        - InstantStyle : Free Lunch towards Style-Preserving in Text-to-Image Generation
    - Personalized generation for multiple concepts
        - Multi-Concept Customization of Text-to-Image Diffusion
        - Break-A-Scene: Extracting Multiple Concepts from a Single Image
- From the lectures on "Vision-Language Generative Models"
    - Video-language
        - CogVLM2: Visual Language Models for Image and Video Understanding
        - LLaVA-OneVision: Easy Visual Task Transfer
    - Multiple input and output modalities
        - NExT-GPT: Any-to-Any Multimodal Large Language Model
        - CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation
    - Variable or high resolution images
        - LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images
        - Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution (they address topics other than resolution, but focus on just the resolution part)
- Other setting could include: 3D voxel, video-audio, 3D meshes, human keypoints

Suggested topics & papers for "Diving into a biomedical application":

- From the lectures on "Vision Diffusion and Generative Models"

    - CT denoising
        - CoreDiff: Contextual Error-Modulated Generalized Diffusion Model for Low-Dose CT Denoising and Generalization
        - Diffusion Probabilistic Priors for Zero-Shot Low-Dose CT Image Denoising
    - Protein structure from Cryo-EM imaging (they use generative models, but not diffusion models)
        - Reconstructing continuous distributions of 3D protein structure from cryo-EM images
        - CryoGAN: A New Reconstruction Paradigm for Single-Particle Cryo-EM Via Deep Adversarial Learning
    - Diffusion for anomaly detection
        - AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise
        - Diffusion Models for Medical Anomaly Detection
    - Ambiguous / uncertain medical segmentation
        - Ambiguous Medical Image Segmentation using Diffusion Models
        - Stochastic Segmentation with Conditional Categorical Diffusion Models
    - You can find more suggestions on this review site
- From the lecture on “Vision-Language Generative Models”
    - General biomedical vision+language
        - BiomedGPT: A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks
        - Med-Flamingo: a Multimodal Medical Few-shot Learner
        - HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale
        - MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models
    - Radiology vision-language model
        - Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation
        - Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data (warning: this paper is a bit longer than normal)
    - Vision-language systems for other targeted clinical applications
        - Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4
        - Integrated image-based deep learning and language models for primary diabetes care