Applied Machine Learning Final Project Proposal

Matthew Liu and Luca Williams

Introduction / Motivation

- Broad Question
    - How do AI models deal with conceptually inferring unseen data (out-of-distribution generalization)?
- Narrower Question
    - How might different attributes of a prompt affect the performance of a model on out-of-distribution generalization?
- Outcomes
    - We are looking to quantify attribute performance in the task of OOD generalization. It would be interesting to see if some attributes are more important than others, or if some combination of attributes worked specifically well together.

Background / Literature Review

1. What were the goals of that paper? How is it related to your project?
2. What methods did the paper use?
3. What were the conclusions?

- Meta-Prompting for Automating Zero-shot Visual Recognition with LLMs
    - Goals
        - The goals of this paper are to answer the question: Does manual design of LLM prompts bias the resulting VLM prompts?
        - They assumed that by letting the LLM produce the text prompt with minimal human guidance, VLM performance would be better.
    - Relevance
        - This paper could help us extrapolate and regulate the prompts we use in the testing of different prompts and attributes.
        - It would give us a larger dataset of prompts to work with and allow control over the attributes that it produces and tunes.
    - Methods
        - MPVR has a two-stage pipeline in how it works.
        - Compose a meta prompt with a system prompt, an in-context example, and a downstream task specification.
        - The LLM responds with a diverse set of generic query prompts, and the class name is then inserted.
        - These prompts are then sent to the VLM.

        - The methods here don't seem too relevant, as this may be more of a tool that we use to generate prompts rather than borrow from its methodology.
    - Conclusions
        - The model consistently outperforms the CLIP zero-shot baseline from other prompts generated by commercial LLMs.
        - Prompt diversity matters → important consideration with attributes.

- ArGue: Attribute-Guided Prompt Tuning for Vision-Language Models
    - Goals
        - The goal of this paper is to use visual attributes to encourage models to use correct rationales for identifying objects; e.g., the object in the sky is a bird because it looks like a bird, not because it is in the sky.
        - The methods proposed are aimed at improving transfer learning, meaning that a model trained for one thing can be used for another.
    - Relevance
        - This paper is directly relevant to our paper, as it studies how a focus on a certain attribute can enable more OOD generalization. It looks at one relevant attribute and is able to show that that attribute is relevant.
        - This is great for our paper because it shows that certain attributes may have a greater impact on OOD performance, and if we can extrapolate some of these methods to other attributes, it may enable us to compare and quantify their efficacy.
    - Methods
        - The ArGue method encourages training a model on confidence in associated visual attributes, so that attribute-based prompts will be more effective than simply using classes.
        - Attribute sampling makes the method more efficient and accurate, and is sampled based on certain criteria.
        - Negative prompting to test if attributes with no connection to classes also demonstrate similar results.
        - Prompt regularization to make sure that the prompt does not overfit and stays close to natural text.
    - Conclusions
        - Negative prompting helps the model not rely on random correlations by enforcing uniform predictions for attributes that have no significant meaning.
        - Attribute sampling improves quality.
        - ArGue outperformed baselines in OOD generalization.

- Align Your Prompts: Test-Time Prompting with Distribution Alignment for Zero-Shot Generalization
    - Goals
        - Highlight problems in models such as CLIP with how it handles OOD
        - Improve the zero-shot generalization of vision-language models when the data is out of distribution

        - In layman's terms, they argue that an aspect of prompt tuning that is overlooked is that it is successful because it tends to adapt the prompt to the data that the model has been trained on, not vice versa.
        - The goal of the paper is to propose the prompt tuning method along with the distribution alignment to keep it in line with the test data, with their method PromptAlign.
    - Relevance
        - The relevance to our project is that it provides a conceptual framework for understanding how prompt engineering interacts with the success of OOD generalization.
        - It helps us understand that the success of a prompt likely depends on the distribution of the data.
        - The evaluation metrics in the paper may come in useful when setting up our experimental design, such as zero-shot evaluation.
    - Methods
        - Start with CLIP as a base model
        - Add learnable tokens to the base model inputs (time test prompting)
        - For each test, update prompts to try and increase confidence ratings
        - Alignment method to align tokens closer to the source dataset.
        - Evaluation compared to other models in zero-shot
    - Conclusions
        - The paper concludes that their method, PromptAlign, demonstrates much better performance than other zero-shot generalization methods in domain generalization and cross-dataset evaluation settings.

- Some other potentially useful preprints (not part of the literature review but still valuable)
    - Title: GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation
        - Reference: arXiv:2406.13743
        - Goals
            - Build a benchmark (GenAI-Bench) to test whether generative image and video models can handle compositional combinations in prompts.
            - Measure how well current models follow complex instructions.
            - Evaluate whether current automated metrics actually reflect human correctness.
            - Propose VQAScore, a metric that better matches human judgment and can improve generation quality.
        - Methods
            - Collected ~1,600 real-world prompts requiring compositional reasoning (relationships, attributes, negation, logic, counting, etc.).
            - Generated images/videos using state-of-the-art models.
            - Conducted large-scale human evaluation (~80,000 ratings).
            - Compared common evaluation metrics (CLIPScore, PickScore, ImageReward, etc.) to human judgments.

        - Introduced *VQAScore*, where a VQA model answers questions derived from the prompt.
        - Tested image-ranking interventions using VQAScore.
    - Conclusion
        - Even strong models consistently fail on complex or unseen compositional structures.
        - Failure varies by category: attribute binding, spatial relations, counting, and negation are especially weak.
        - Existing automated metrics correlate poorly with human correctness for compositional/unseen prompt categories.
        - VQAScore correlates much better with humans and improves accuracy when used to rank multiple generations.
        - Compositional reasoning remains a major limitation in current generative models.

    This paper introduces a benchmark designed to evaluate whether generative models can reliably handle compositional prompts. Using a dataset of approximately 1,600 real-world prompts requiring complex relationships, attributes, and logical structure, the authors assess images produced by state-of-the-art models and find that even the strongest systems consistently fail on many compositional tasks. The study further examines whether common automated metrics (CLIPScore, PickScore, and ImageReward) accurately reflect human judgments. It concludes that these metrics correlate poorly with human evaluations when compositional reasoning is involved. To address this gap, the authors introduce VQAScore, a VQA-based alignment metric, and show that ranking generations with VQAScore yields substantially better agreement with human judgment. Overall, the paper demonstrates that compositional reasoning remains a significant limitation for current generative AI models.

- Title: GenAI Confessions: Black-box Membership Inference for Generative Image Models
    - Reference: Matyas Bohacek, Hany Farid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2025, pp. 321-330
    - Goals
        - Provide a method for membership inference (i.e., decide if image xxx was in the training set of a generator) for text-to-image/diffusion models.
        - Demonstrate the feasibility of such attacks under limited (black-box) conditions, which have implications for privacy, copyright, data provenance, and trustworthiness of generative models.
    - Methods
        - The authors assume black-box access: you can query the generative model, but you do *not* have access to its weights or internal representations.

        - They develop a membership inference attack that compares generated outputs (via querying the model) and uses similarity or statistical measures (between suspected member image and generated images) to infer membership.
        - They perform experiments on generative image models (including diffusion models or image-to-image pipelines) using query images and measuring success (ROC-AUC, etc.) to validate that the attack works.
        - They also analyze how various parameters affect success: query access, auxiliary data, shadow models, size of training set, etc.
    - Conclusion
        - The method demonstrates that even under black-box conditions, it is possible to infer whether an image was part of a generative model's training set, with non-trivial accuracy.
        - This raises privacy, intellectual-property, and trustworthiness concerns for generative image models: if a model has memorized and can leak training data membership, then the confidentiality of training datasets is at risk.
        - The findings suggest the need for auditing tools, dataset provenance checks, and possibly mechanisms to defend against membership inference in generative models.

    The paper addresses the question of whether we can determine if a specific image was used to train a generative image model, under black-box access (i.e., without knowledge of the model's internals). The authors propose a membership-inference attack that queries a generative model and compares the outputs to a target image, showing that diffusion-based image generators reveal measurable differences between images they have memorized and those they have not. Through experiments across multiple datasets and model configurations, the paper demonstrates that generative models are vulnerable to this type of inference, achieving non-trivial membership-prediction accuracy without any access to model internals. These results reveal significant privacy and copyright risks in modern generative systems and underscore the need for stronger safeguards, auditing tools, and transparency around training data. By applying their method, we can empirically separate candidate test images into "likely seen" and "likely unseen" categories. This allows us to construct a more reliable unseen-data benchmark and evaluate how well generative models handle images or concepts not present in their training distribution.

Planned Methods

- **Define OOD tasks.**
    - We first need to have some sort of task that we are confident is OOD to have the model perform, most likely with image generation.
    - There are two ways to determine this data:

        - Google images as a proxy: if that image concept is sparse in the Google images page, then it is likely that the model might not have been trained on it.
        - Method from the preprint above to use membership inference to determine whether the model has seen it or not.
- **Establish differences in attributes to focus on and prompts to use**
    - Could use the MPVR model above to automate prompts created by LLMs (I think that is what it does, but not 100% sure)
    - We would establish a list of attributes to look at and then implement them into prompts based on the OOD tasks above. Different prompts would have different kinds and amounts of attributes.
- **Collect output data on model performance.**
    - We could use evaluation metrics such as VQAScore, as described in a preprint above, to measure the success of a model, prompt, and OOD task.

Proposed Timeline

Task	**Time required**	**Expected date of completion**	**Person**
List of attributes	1 day	Thursday, November 13	Both
OOD Tasks	3 days	Monday, November 17	Luca
Adapt VQAScore	3 days	Monday, November 17	Matthew
MPVR Model	2 days	Wednesday, November 19	Both
Milestone 1: Have a demonstration of the VQAScore, prompt generation, and indication of the difference in attributes.	2 days	Friday, November 21	Both
TBD*			
Milestone 2: Poster Presentation		Thursday, December 4	Both

* We would wait to see how the work towards milestone 1 goes, then reassess and map out the rest of the work. If all goes well before milestone 1, the majority of this work will be expanding our methods and data collection to more OODs and more attributes to be able to generalize trends.