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

Week 11: Dataset and Model Biases

Instructors: L.-P. Morency, A. Zadeh, P. Liang *Synopsis Leads: David Lin, Yuchen Xu*

Edited by Paul Liang *Scribes: Chonghan Chen, Kelly Shi*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

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 11's discussion session, the class aimed to formalize a taxonomy of bias in terms of sources, causes and effects. Students identified challenges in bias detection and proposed possible solutions to mitigate bias. Future works were suggested in curating high-quality dataset, designing fairness-aware model architectures, and improving model interpretability.

The following list of research probes was provided:

1. What is a taxonomy of biases in multimodal datasets and models?
2. What are some risks related to biases (e.g., social biases) when creating new datasets? How are these risks potentially amplified or reduced when the dataset is multimodal, with heterogeneous modalities? Are there any biases that are specific to multimodal data?
3. What are the imperfections that may arise during human annotations? How do these imperfections in data and labels affect multimodal learning of multimodal representations, cross-modal interactions, co-learning, and pre-training?
4. Can biases also emerge not only from the multimodal training data, but also from the modeling design decisions themselves? What aspects of multimodal modeling are most prone to learning and possibly emphasizing biases?
5. What are potential solutions for tackling these risks and biases in multimodal datasets and models? How can we properly identify, visualize and eventually reduce these biases in multimodal datasets and models?
6. How can we better interpret multimodal datasets and models to check for potential biases? What specific dimensions should we strive to understand?
7. What are the tradeoffs between large-scale, noisily-collected and annotated multimodal datasets versus small-scale, carefully-curated and annotated datasets? How do these affect multimodal modeling? How does it relate to the popular pre-training paradigm?

As background, students read the following papers:

1. (Required) Measuring Social Biases in Grounded Vision and Language Embeddings [Ross et al., 2021]
2. (Required) Shortcut Learning in Deep Neural Networks [Geirhos et al., 2020]
3. (Suggested) Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes [Birhane et al., 2021]
4. (Suggested) A Case Study of the Shortcut Effects in Visual Commonsense Reasoning [Ye and Kovashka, 2021]
5. (Suggested) Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets [Geva et al., 2019]
6. (Suggested) DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative

Transformers [Cho et al., 2022]
7. (Suggested) Revisiting Visual Question Answering Baselines [Jabri et al., 2016]
8. (Suggested) Analyzing the Behavior of Visual Question Answering Models [Agrawal et al., 2016]
9. (Suggested) Adversarial Filters of Dataset Biases [Le Bras et al., 2020]
10. (Suggested) Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics [Swayamdipta et al., 2020]
11. (Suggested) Annotation Artifacts in Natural Language Inference Data [Gururangan et al., 2018]
12. (Suggested) AI and the Everything in the Whole Wide World Benchmark [Raji et al., 2021]
13. (Suggested) Perceptual Score: What Data Modalities Does Your Model Perceive? [Gat et al., 2021]
14. (Suggested) Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers [Peng et al., 2021]
15. (Suggested) Challenges in Real-life Emotion Annotation and Machine Learning Based Detection [Devillers et al., 2005]

We summarize several main takeaway messages from group discussions below:

1 A Taxonomy of Biases

1.1 Taxonomy

We can divide biases into two categories according to their source: **dataset bias** and **modeling bias**. Examples of dataset biases include uneven number of samples for each category, correlation between a category and an irrelevant attribute, and data distribution that reflect social stereotypes. An example for modeling bias is that CNN-based models inherently rely on texture instead of shape and other local information. Some ways to resolve dataset biases include data augmentation. To reduce modeling bias, we could try adding special losses enforcing some extra constraints.

Biases can be intrinsic to certain tasks. For example, different languages may each contain some language-specific biases, and datasets constructed with that language will inevitably contain such biases.

We can also divide biases into **shortcut bias** and **fairness bias**. Shortcut biases make models rely on superficial signals to make predictions, which harms generalization to out-of-domain data distributions [Geirhos et al., 2020]. Fairness biases tend to exacerbate social stereotypes and compromise equality when a model trained with such biases is in production.

1.2 Not All Biases Are Equal

Some biases result from learning with a dataset that has dataset bias. For example, an image classifier may learn to predict an object by only looking at the background when the object always co-occur with the background in the training set. Other biases may be invariant to the dataset, and are model-specific.

We also observe that what we consider as fairness biases can be useful in some scenarios. For example, in the healthcare systems, information such as gender, race and age may be important for a model to make a prediction, but users then have to be careful when deploying and visualizing these models.

2 Biases in Multimodal Problems

2.1 Adding a New Modality

If we have an unbiased modality or a modality contradicting the bias held by other modalities, there are two possible outcomes: (1) The bias gets mitigated. More specifically, the model learns the relations with professions without being affected by the biases in other modalities. (2) The biases in other modalities are amplified. The model may learn to exploit biases in other modalities. Ideally, we would like the model to debias itself with the contradictory modality. However, there is no evidence available to show how we could achieve this goal.

In fact, having an unbiased modality might amplify the bias, as the current systems are not specifically designed to recognize relations among biases in multiple modalities. Our current systems assume that modalities are correlated (because of alignment), therefore if one modality is biased, all of them will be biased. We should design a good multimodal system that models biases and stereotypes therefore we could mitigate biases if there exists an unbiased modality.

There may be other factors that affect the bias. For example, if we have one modality that contradicts the stereotype, but this conflict may not be picked up by the model. Their model might be learning some shortcut rather than learning the conflict. In this case, the unbiased modality might be interpreted by the model in a way that adds to the stereotype.

Besides examining biases represented by the data, we also need to check if the model is greatly biased to certain modalities. For example, if the model is dominated by language, then any bias in other modalities or any features that could correct biases in language will not have much effect on model output. In order to mitigate biases, we need to be able to understand how the model makes decisions.

2.2 Dominant Modality

In multimodal machine learning, we often face the problem of dominant modality. The dominant modality has excessive contribution to model predictions as well as to model biases. The language modality, in most cases, is the dominant modality in prediction. Since the biases represented by the dominant modality are more likely to the stronger biases learned by the model, we could design models that balance the contribution of each modality.

3 Detecting Biases

Biases are often dataset dependent. For the language modality, word co-occurrences are often good indicators of biases. For example, gender-related words often co-occur with certain profession-related words. In this case, we can perform masking experiments on deep language models such as BERT to verify how much they rely on such information.

Adversarial data augmentation can be used to test models' robustness. One example of this is to append shortcut information to each image in the dataset. If the model is not robust enough to ignore the shortcut information during training, it will end up with poor test set performance. The counterpart for the language modality is appending the answer to each question in QA datasets. A more sophisticated way is to combine DGGAN [Chen et al., 2020] with an RL-based algorithm to generate harder and harder samples that target the weaknesses of models. We can also inspect the generated examples and get intuitions on what the weaknesses are and what biases may exist.

4 Mitigating Biases

We propose several possible approaches to mitigate biases in the multimodal setting.

4.1 Data-centric Approaches

Several techniques used to mitigate dataset biases involve either perturbing or augmenting data.

4.1.1 Dataset Curation

To avoid bias, we should collect data with minimum bias and curate high-quality datasets. Peng et al. [2021] show that dataset retraction has a limited effect on mitigating harms. The underlying data remained widely available, so retractions are unlikely to cut off data access. To collect a large dataset without introducing too much bias, we could first build a small dataset that is built specifically to classify the level of bias represented. Then we can use this classifier to filter out data with high bias.

4.1.2 Data Perturbation

To reduce bias in image modality, we could separate the subjects and backgrounds using image segmentation, then substitute the subject or the background to achieve a balanced dataset and reduce the chance of shortcut learning. We could also obfuscate or replace certain words that cause biases more easily. However, there may be a trade-off between masking bias-prone words and model performance. In certain cases, gender information is crucial for making correct predictions.

4.1.3 Data Augmentation

Domain randomization and RL-based algorithms can be used to augment the data. For example, Ramaswamy et al. [2021] use GANs to generate realistic-looking images, and perturb these images in the underlying latent space to generate training data so that the data is balanced among protected features such as race and gender.

4.2 Input Features

Choosing proper features could be one potential solution to mitigate biases. For example, instead of feeding images, we could feed facial movement data leading to a better-trained model that is less biased. However, there is a trade-off between feature engineering and learning well-informed data representations. We could also focus on features that are prone to representing biases and manually engineer these features to mitigate biases.

4.3 Model Design

We could also design fairness-aware models. FairGAN [Xu et al., 2018] is a fairness-aware GAN that accepts a list of protected attributes as input. Whenever the discriminator sees a bias towards the protected attributes, the model will be penalized. However, biases are not just in terms of gender or race, correlations between any two entities could be biases. Instead, we could also add a fairness term to the loss function, so that the model also optimizes to achieve a more fair prediction.

5 Reasoning and Interpretability

Interpretability helps detect and identify biases because decision rules of models can be examined if they are made explicit. For example, we can check whether a model focuses on irrelevant features or a highly biased modality to make a prediction. One way to enforce interpretability and eliminate biases is to force models to do “reasoning” in the same way as we humans do it. To achieve this, we could define a “gold” reasoning path and impose losses whenever the model fails to follow. If we define a perfect imitation as ‘unbiased’, then we could apply imitation learning techniques to help debiasing. For example, we could let a model observe human brain signals through fMRI and learn from that. However, it is unclear whether the ultimate goal for debiasing is to imitate human thought processes. We also discussed other concerns about this approach.

- This approach will require extra annotations because we supervise not only with the label, but also with the reasoning path. This is sometimes not affordable.
- Models may take shortcuts to minimize the path-following loss.
- It is difficult to define how humans do a task. In other words, the “gold” reasoning path itself can be ambiguous.
- There may be more than one valid reasoning path. It is unclear how to train a model to learn a good reasoning path when there are multiple of them.
- Human thought processes can be biased to begin with, and imitating human thought processes only replaces a set of biases with another.
- It is ambiguous about how we break down a task into separate reasoning steps a model should take along the reasoning path.

6 General Challenges

Detecting biases with special tests may not always expose a model's bias. For one thing, it is unlikely that a test exhaustively targets all possible biases. For another, when we use a benchmark to evaluate how robust a model is against biases, even when we observe an increase in scores, we cannot tell whether a model only hides the biases that are evaluated by the benchmark as opposed to truly improving the estimated distribution by discarding biases. For example, an image generation model may be tuned to get high benchmark scores by generating males and females with equal likelihoods given certain occupations. However, when the model is given a gender and asked to generate an image of this individual at work, biases may still exist, which are not evaluated by the benchmark.

Another issue is that bias can be subjective. For example, some memes are only entertaining when a strong bias is included. Even when they are created and used by a group of people who do not consider them as hateful at all, they could be offensive to people from a different culture. One could argue these biases should not be classified as harmful because the memes were not created for hateful purposes. Others may think that the biases should be eliminated with the memes as they bring a negative impact to some people.

We face even more challenges when multiple modalities are involved. Some biases may only appear when two or more modalities are combined. Any detection and debiasing approach will be required to model cross-modal interactions to deal with these cases.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, 2016.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, Wei Fan, and Xiaohui Xie. Dggan: Depth-image guided generative adversarial networks for disentangling rgb and depth images in 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 411–419, 2020.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, 2019.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112, 2018.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR, 2020.

Kenneth L Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9301–9310, 2021.

Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, 2021.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, 2020.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.

Keren Ye and Adriana Kovashka. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3181–3189, 2021.