COMP 472 Artificial Intelligence (Summer 2022)

Project Assignment, Part II

Due date (Moodle Submission): Wednesday, June 22
Counts for 50% of the course project

AI Face Mask Detector: Update. The goal of this second part of the project is to perform an evaluation for a possible *bias* of your AI and (at least partially) eliminate it. Additionally, you have to fix any issues that came up during the demo of the first part.

Bias in AI. With an increasing number of AI-based applications being deployed in practice, the analysis of a possible *bias* – introduced through the training data – has become a major concern.[1] In this project part, your goal is to analyze your AI for **two** of the following categories:[2] *age, race,* or *gender.* For example, based on suitably annotated testing data, you can check if your performance for the four classes (cloth mask, surgical mask, FFP2/(K)N95 mask, no mask) is consistent for male and female faces.[3] To eliminate the bias, experiment with re-balancing and enhancing your training dataset. Evaluate the performance of your network both on the complete dataset and the chosen bias attribute subset(s).

Evaluation: K-fold cross-validation. In addition to the basic train/test split from Part I, you have to improve your evaluation across the different classes using *k-fold cross-validation.*[4] Perform a 10-fold cross-validation evaluation (with random shuffling) on your AI and add the results to your report. Make use of *scikit-learn*/skorch for splitting your datasets (i.e., do *not* use a manual, static split).[5]

Report. Update your report, correcting any issues from the first version and adding a new chapter for the bias detection & elimination, as well as new evaluation sections for the cross-validation. That is, provide a single, combined report that includes the sections from Part I (updated for any changes to your model) and the new sections below:

K-fold cross-validation: Add the results of your 10-fold cross-validation for *both* the original model (that you saved in Part I of the project) *and* your re-trained final model (after the bias analysis and re-training), with the following information for each model:

1. Provide a table that shows, for *each* of the 10 folds, the precision, recall, F_1-measure, and the accuracy.
2. Then, provide the aggregate statistics across the 10 folds, again for precision, recall, F1-measure, and accuracy.
3. Compare the results you obtained for the original model (from Part I) using cross-validation with your previous results, using the single training/test split. Do you get some additional insight from the cross-validation?

Provide any insights regarding your system's performance in an analysis section (i.e., explain differences between k-fold evaluation and your original train/test split evaluation).
Length: ca. 3/4 page for the analysis, plus tables

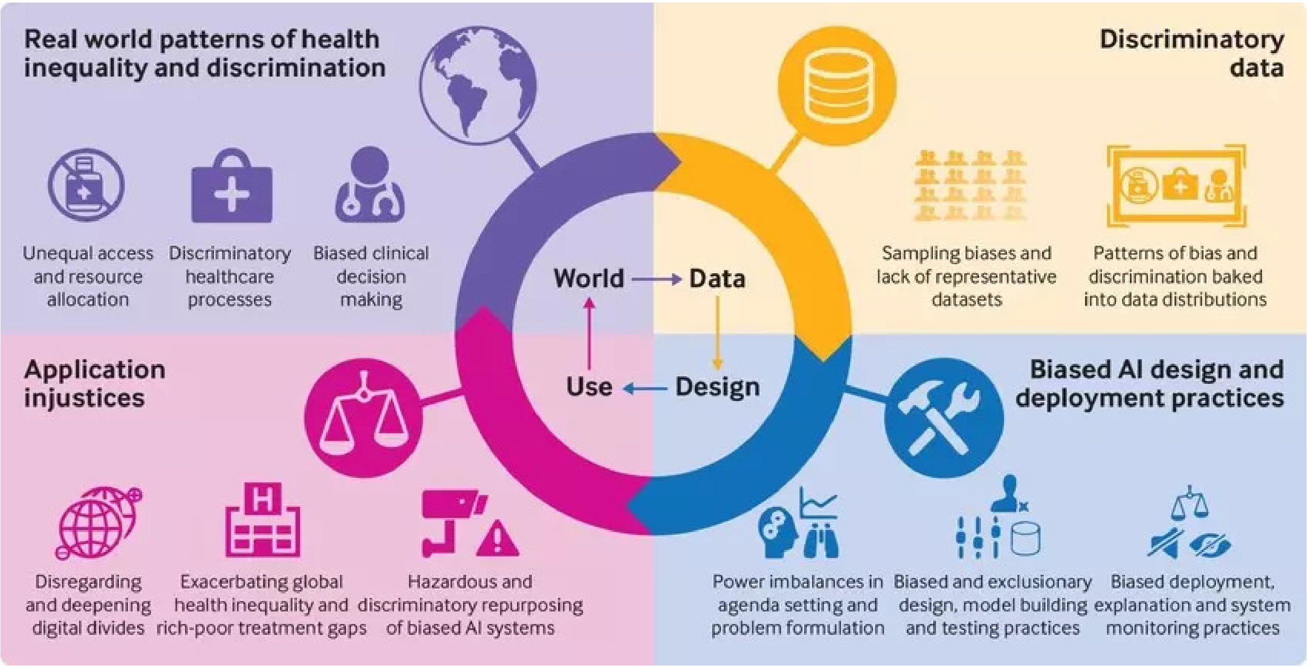
[1] E.g., *"Amazon scraps secret AI recruiting tool that showed bias against women"*, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[2] If your group has only two members, you have to analyze only one category

[3] See https://www.media.mit.edu/projects/gender-shades/overview/ as an example for systematic AI product testing.

[4] See https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-iterators

[5] See https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html


Real world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World
Data
Use
Design
Application injustices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Biased AI design and deployment practices
Power imbalances in agenda setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices

Bias in AI (World Economic Forum[6])

Bias: Describe which of the above attributes you are analyzing for bias and explain your process for bias detection. Provide the following evaluation details for both the original model (from Part I) and your new, re-trained model:

- Performance evaluation metrics for each of the subclasses (e.g., different age groups) you analyzed for bias (precision, recall, F_1-measure, accuracy).
- Confusion matrices for each subclass and a confusion matrix for the whole dataset (make sure the classes are always clearly labeled).

Describe how the bias was introduced in your AI, how you addressed and re-trained it and compare the performance between the model from Part I and your new model.
Length: ca. 1 pages, plus tables

Deliverables. The structure of the deliverables is the same as for Part I (with your updated resources and report, of course). Submit the complete project, not just the changes you did for Part II:

Python code: All the Python code that you developed for this project. You must have a complete CNN implemented using PyTorch. You must submit executable Python programs; notebook files (e.g., Colab, Jupyter) are not accepted.

Dataset: Information on the datasets you collected, as well as a file detailing the source of each dataset/image. For external, publicly available dataset, only include a reference to the source with the details mentioned above. Include images you created yourself, as well as any manually created metadata for the Phase II bias evaluation.

Trained Models: The two trained models that you used in your evaluations (the one from Part 1 of the project and the new one), together with some sample data (ca. 100 images) and instructions on how to run your system on the provided data.

[6] https://www.weforum.org/agenda/2021/07/ai-machine-learning-bias-discrimination/

README: A `readme.txt` (or `readme.md`) file that lists all submitted files with an explanation of their content. It also must describe how to run your code for (a) training and (b) testing (including generating the evaluation results provided in the report). If your instructions are incomplete and your code cannot be run you might not receive any marks for your work.

Report: The project report, as detailed above, as a PDF.

Submission. Submit the updated version of your project through Moodle by the due date (late submission will incur a penalty, see Moodle for details). *If your group's members have changed from Part #1:* Include a single, signed by all team members, *Expectation of originality* form (see https://www.concordia.ca/ginacody/students/academic-services/expectation-of-originality.html) with your submission (this can be the same form as for Part I if your team did not change).

Demo. We will schedule demo sessions for your project using the Moodle scheduler. Demos will take place on-campus, in-person. All team members must be present for the demo.