ICCV23
PARIS

ECML PKDD 2024

Mining bias-target alignment from Voronoi Cells

ECML-PKDD 2024 Nectar Track presentation
10/09/2024

TELECOM Paris

IP PARIS

Remi Nahon, Van-Tam Nguyen, Enzo Tartaglione
Maître de Conférences, équipe MM, dept.IDS, LTCI

enzo.tartaglione@telecom-paris.fr

Outline

- Introduction: Model debiasing
- Supervised VS Unsupervised debiasing
- Mining bias-Target alignment from Voronoi cells
- Conclusions, current and future research

Bias in AI models

- Model Bias occurs when the model itself is not able to accurately represent the underlying relationship between the input features and the output variable.
- Simply: the model captures some **spurious relations**, harming the performance at test time.
- We can solve this problem providing metadata of these correlations. However, this is an expensive and sometimes is even unfeasible (eg. when these are not known a-priori).

Image taken from https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Algorithmic VS Societal bias

- **Societal Bias** refers to the prejudices, stereotypes, and inequalities that exist in society. These biases are embedded in cultural norms, institutions, and social practices, and can shape the data that algorithms use.

- **Algorithmic Bias** occurs also when these societal biases are encoded into algorithmic systems, either through biased data, biased model design, or biased implementation. While societal biases are often the source, algorithmic bias refers to how these biases manifest in automated systems, potentially amplifying or perpetuating existing inequalities.

Supervised vs Unsupervised debiasing

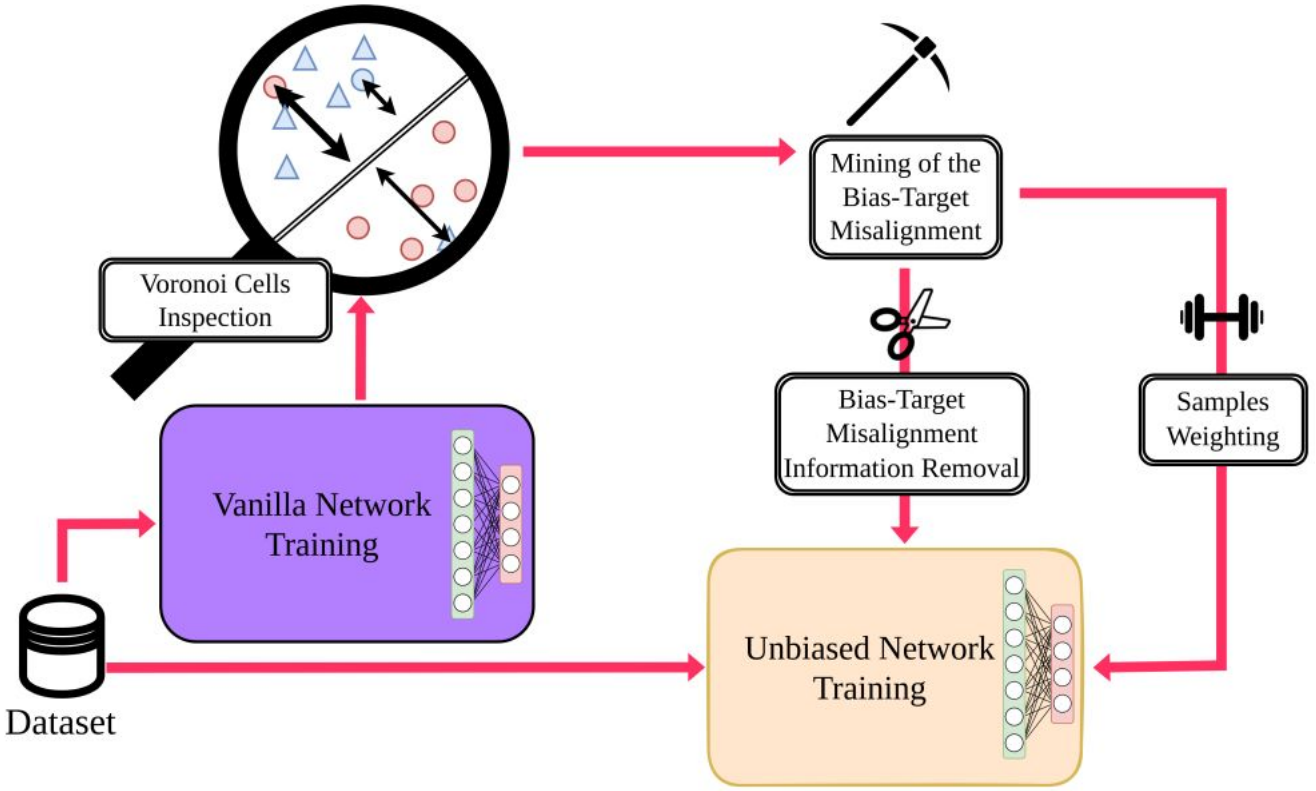
- **Supervised debiasing** refers to approaches that remove the bias when the information related to the bias is already provided.
  - Dataset cleanup approaches
  - Model post-processing
  - In-model approaches (features balance, gradient inversion etc.)

- **Unsupervised debiasing** refers to techniques that identifies and removes the bias without provided information. Some assumptions are always taken:
  - Specific biases are searched for in the dataset (you use a proxy model to find potential biases) – Bias-Tailored approaches (BT)
  - Biases are learned earlier in the training process, and the model fits them better than the target ones (our assumption).

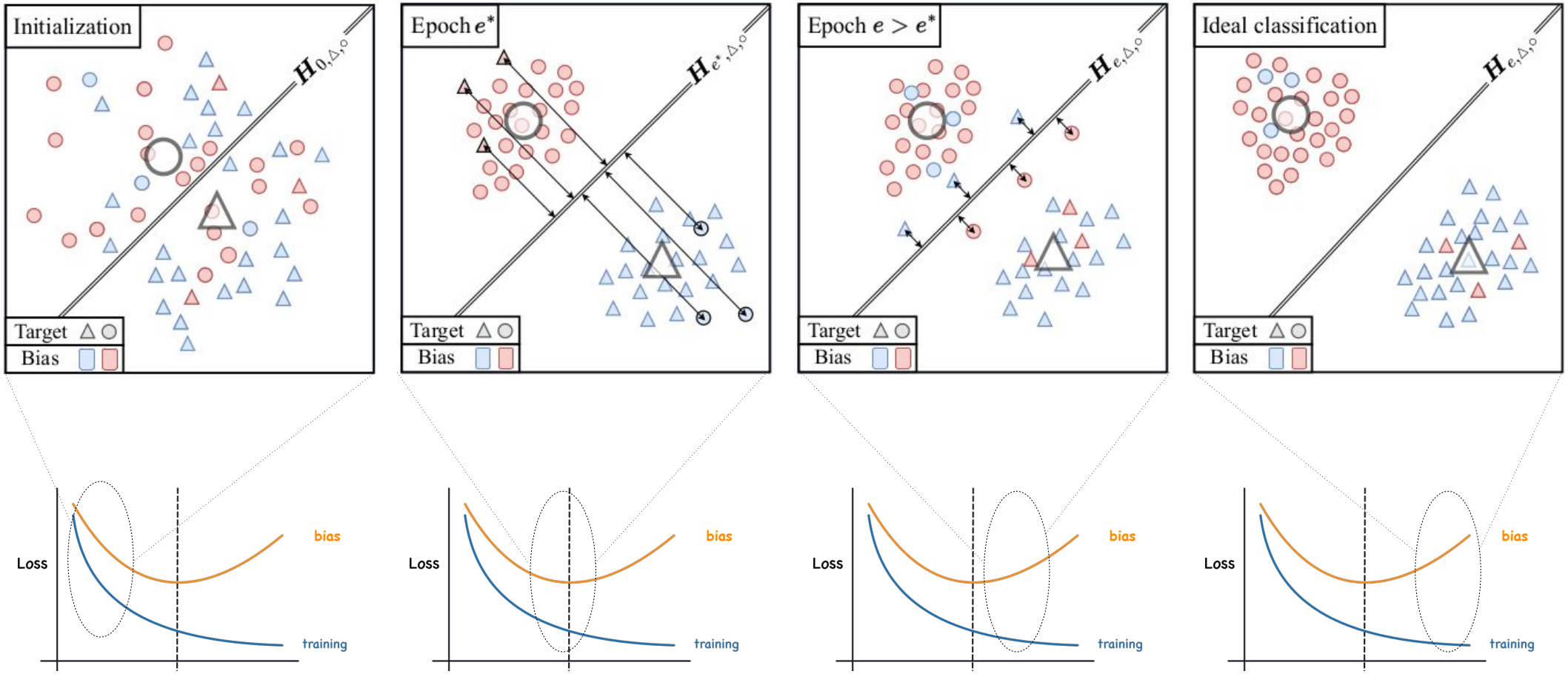
Mining bias-target alignment from Voronoi Cells

- Without knowledge of the bias, we can guess it looking at the distribution of the samples during training.
- The assumption is that the bias is learned before the correct set of features.

- We can achieve comparable performance than being provided with the bias metadata.



Voronoi Cells Inspection



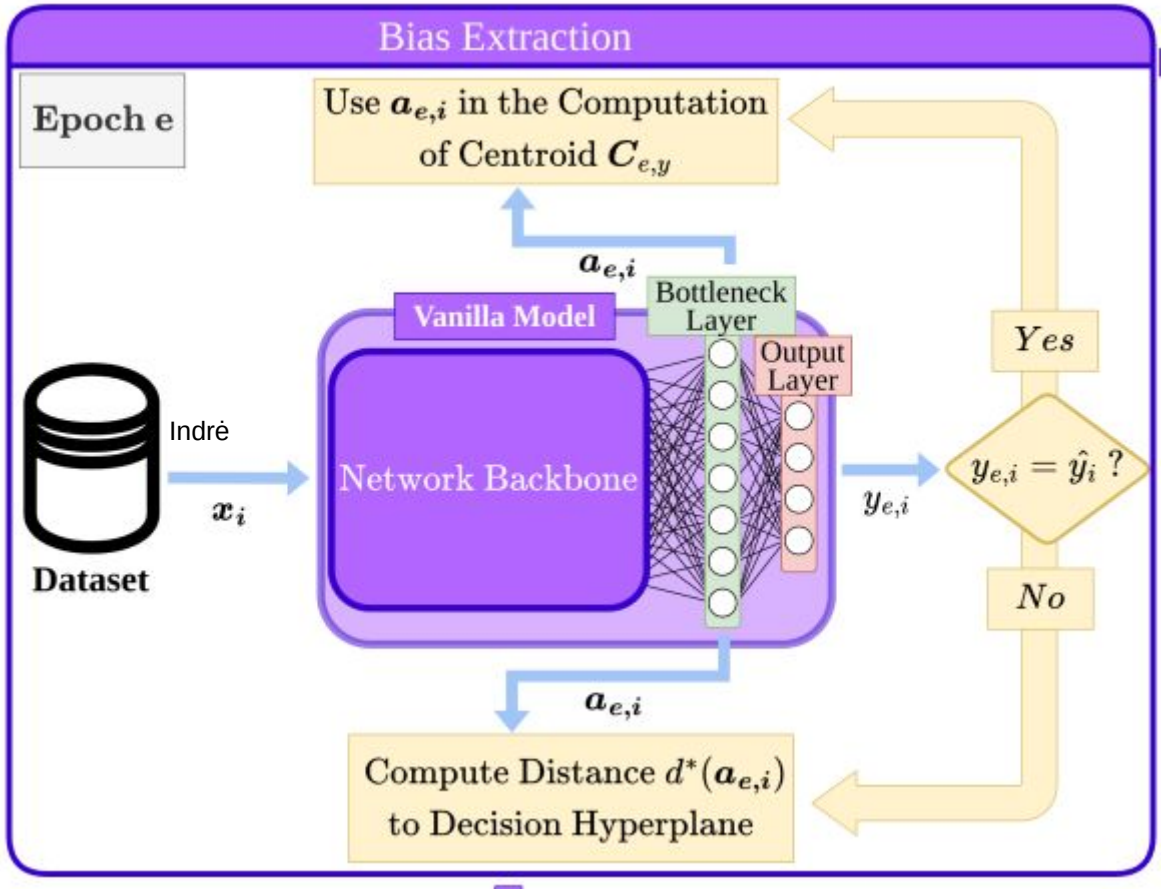
Bias Extraction

- From vanilla training we identify the best epoch e* for the bias extraction looking at

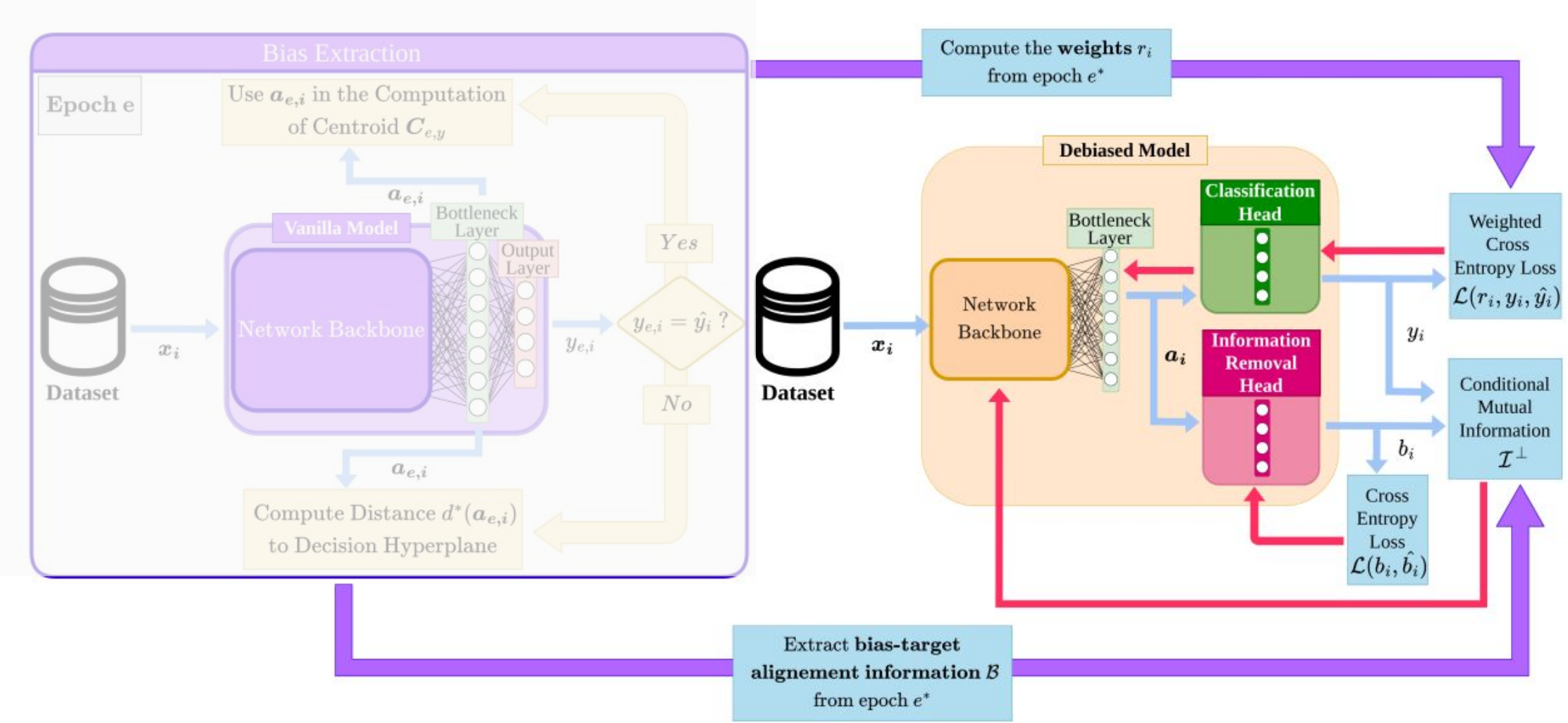
$$e^* = \underset{e}{\operatorname{argmax}} \frac{\sum_i d^*(\boldsymbol{a}_{e,i})}{\|\mathcal{D}_e^{\perp}\|_0}$$

$$d^*(\boldsymbol{a}_{e,i}) = \begin{cases} 0 & \text{if } y_{e,i} = \hat{y}_i \\ \|\boldsymbol{a}_{e,i} - \boldsymbol{H}_{e,\boldsymbol{C}_{e,y_{e,i}},\boldsymbol{C}_{e,\hat{y}_i}}\|_2 & \text{if } y_{e,i} \neq \hat{y}_i \end{cases}$$

- This gives us the psudo-labels that we will use for the debiasing process.



Model Debiasing

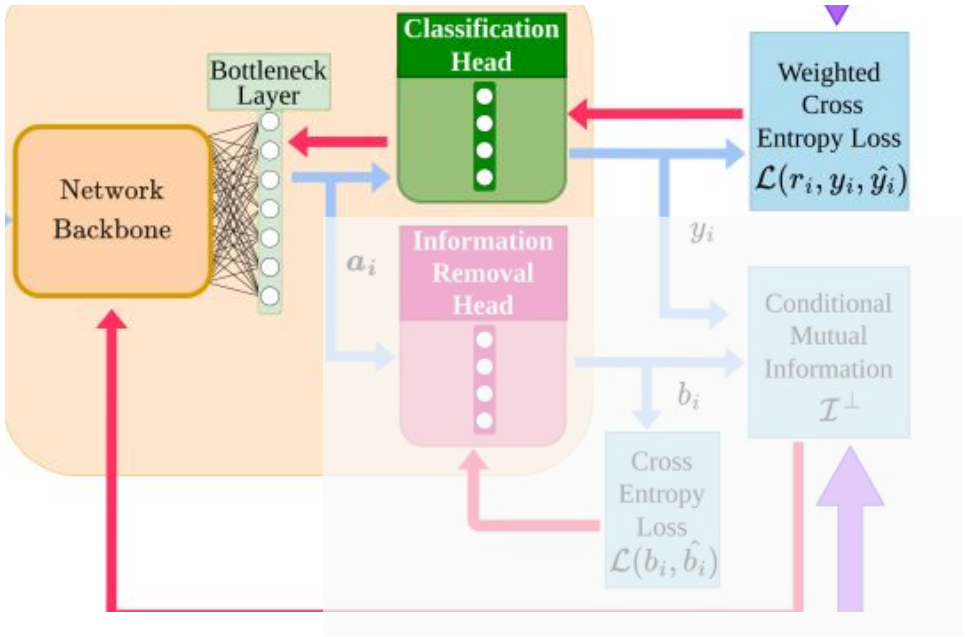


Model Debiasing: Weighted CE

- Here we apply a weight to the loss function relative to the biased subgroups for a target class.

$$r_i = \begin{cases} \dfrac{1}{\rho_{\hat{b}_i}} & \text{if } x_i \in \mathcal{D}^{\parallel} \\ \dfrac{1}{1-\rho_{\hat{b}_i}} & \text{if } x_i \in \mathcal{D}^{\perp} \end{cases}$$

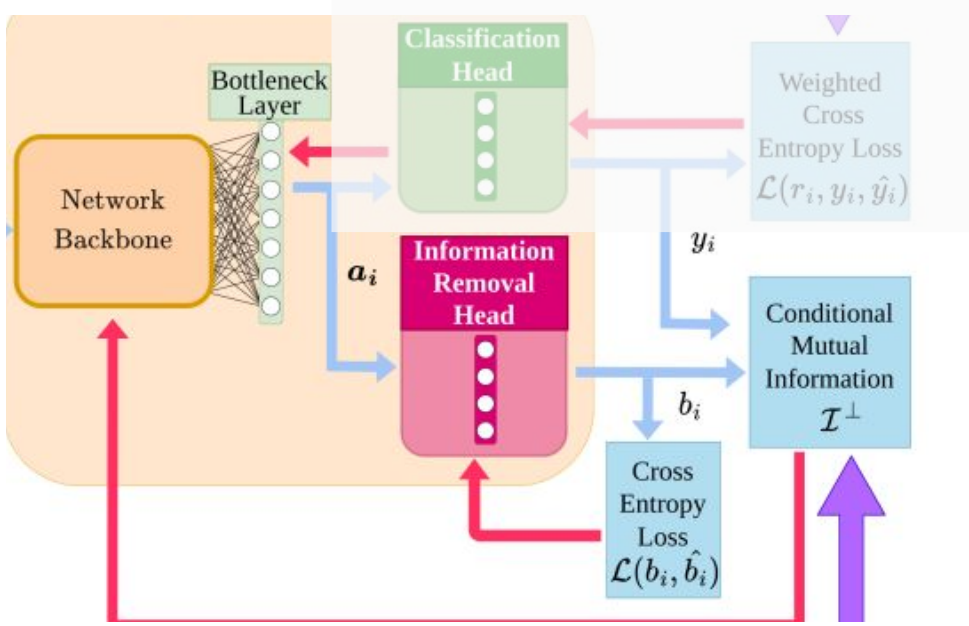
$$\rho_c = \frac{\left\| \mathcal{D}_c^{\parallel} \right\|_0}{\left\| \mathcal{D}_c \right\|_0}.$$



Bias Extraction: Information Removal Head

- The IRH estimates how much information related to the bias can be extracted by the classification head.
- For this it is trained to maximally fit the bias through a cross-entropy loss.

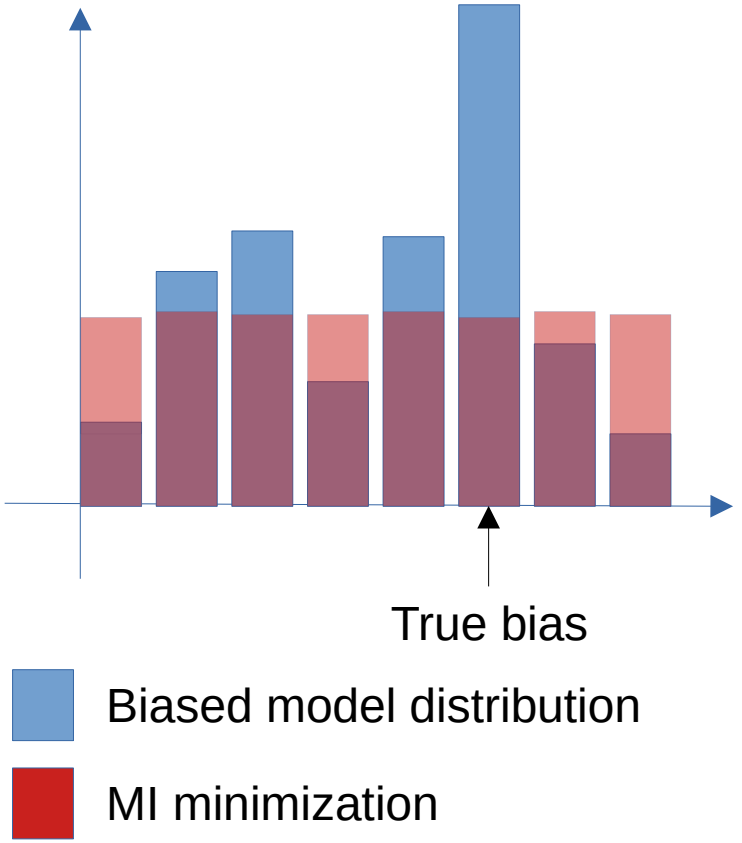
- At the same time we want to discourage the backbone from extracting it at its very root: we minimize the Mutual Information between GT of the bias and the bias itself!



$$\mathcal{I}^{\perp} = \sum_{j,k} p^{\perp}_{b,\hat{b}}(j,k) \log \left[\frac{p^{\perp}_{b,\hat{b}}(j,k)}{p^{\perp}_{b}(j) p^{\perp}_{\hat{b}}(k)} \right] \qquad p^{\perp}_{b,\hat{b}}(j,k) = \frac{\sum_i b_i \cdot \delta_{\text{argmax}(b_i),j} \cdot \delta_{\hat{b}_i,k}}{\|\mathcal{D}^{\perp}\|_0}$$

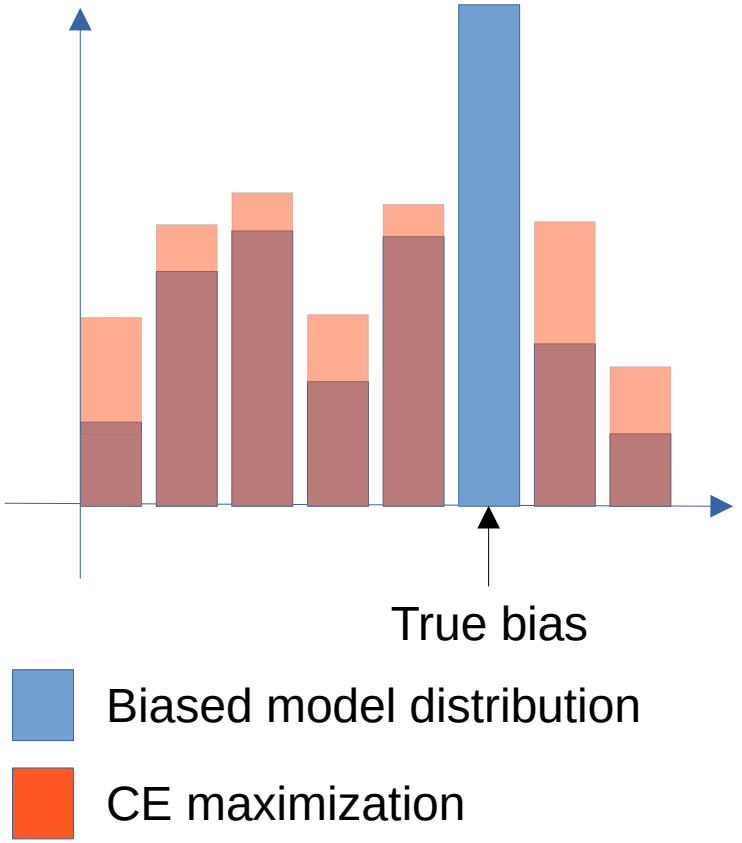
CE maximization vs. MI minimization

- Minimizing the MI makes the class distribution tend to a uniform.



CE maximization vs. MI minimization

- Minimizing the MI makes the class distribution tend to a uniform.
- Maximizing the CE makes the model completely fail for the target class
  - Are we really removing the information?



Results

Table 1: Results on Balanced Biased MNIST when training with different correlations color-digit ρ.

Method	Bias agnostic	Test accuracy [%] (↑) $\rho=0.999$	$\rho=0.997$	$\rho=0.995$	$\rho=0.99$
Vanilla	✓	11.2	40.5	72.4	88.4
Rubi [9]	✗	13.7	90.4	43.0	93.6
EnD [34]	✗	52.3	83.7	93.9	96.0
BCon+BBal [20]	✗	94.0	97.3	97.7	98.1
HEX [37]	BT	10.8	16.6	19.7	24.7
ReBias [4]	BT	26.5	65.8	75.4	88.4
LearnedMixin [12]	✓	12.1	50.2	78.2	88.3
LfF [30]	✓	15.3	63.7	90.3	95.1
SoftCon [20]	✓	**65.0**	88.6	93.1	95.2
Ours	✓	58.7±21.8	**92.7±1.2**	**95.5±0.8**	**97.7±0.4**

Results

Table 4: Test accuracy on 9-class ImageNet and ImageNet-A.

Method	Bias agnostic	Test accuracy [%] (↑)	
		9-class ImageNet	ImageNet-A
Vanilla	✓	94.0	30.5
ReBias [4]	✗	94.0	30.5
StylImageNet [17]	BT	88.4	24.6
LearnedMixin [12]	BT	79.2	19.0
RUBi [9]	BT	93.9	31.0
LfF [30]	BT	91.2	29.4
SoftCon [20]	BT	95.3	34.1
FairKL [6]	BT	95.1	35.7
Ours (BagNet [8])	BT	96.4 ±0.0	34.5 ±3.4
Ours	✓	**95.5 ±0.2**	**34.2 ±0.9**

Beyond this work…

- Paper and code are publicly available!
  See you on Wed at poster stall #45 for more discussion!

- All nice and good… do we know what the bias is?
  - Look at our follow-up!
    “Say My Name: a Model's Bias Discovery Framework”

- It is a two-steps approach – can we do better?
  - Look at our ECCV 2024 paper!
    “Debiasing surgeon: fantastic weights and how to find them”

- Curious to know more about Bias&Fairness in Deep Learning?
  - See you in Milan the 29th for
    FAILED: Fairness and ethics towards transparent AI: facing the chalLEnge through model Debiasing satellite workshop of ECCV 2024.