Information Removal at the Bottleneck in Deep Neural Networks

Enzo Tartaglione
LTCI, Télécom Paris, Institut Polytechnique de Paris

Removing information

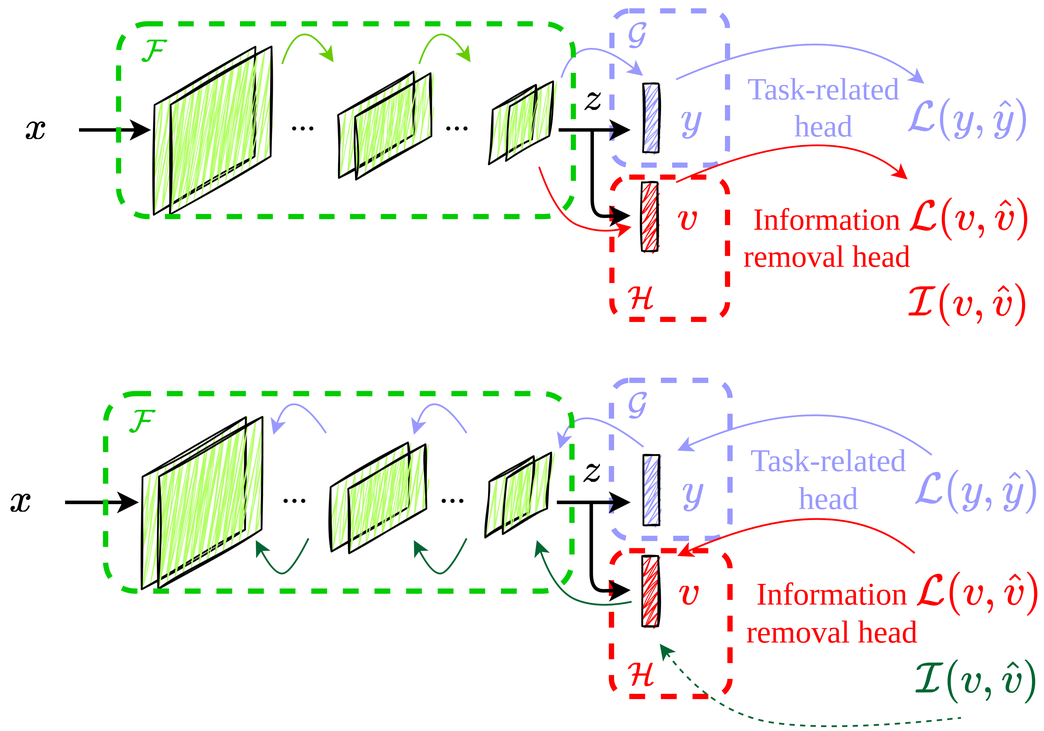
- A lot of interest is recently raised to the themes of fairness and bias in deep learning, against unethical and discriminatory AI.
- Many techniques attempts to tackle these issues, but the metric evaluated is typically accuracy, on balanced datasets: are they really removing information?
- We propose a method to remove specific information at the **bottleneck** of deep neural networks.
- This method relies on the **estimation** of the **information** we desire to maintain **private**, with the employment of an auxiliary classifier.
- Then, we **minimize** a differentiable proxy of the **mutual information**.

IRENE

The input is forward-propagated through the whole network, including the two heads. Losses are computed, along with the differentiable proxy, in the form

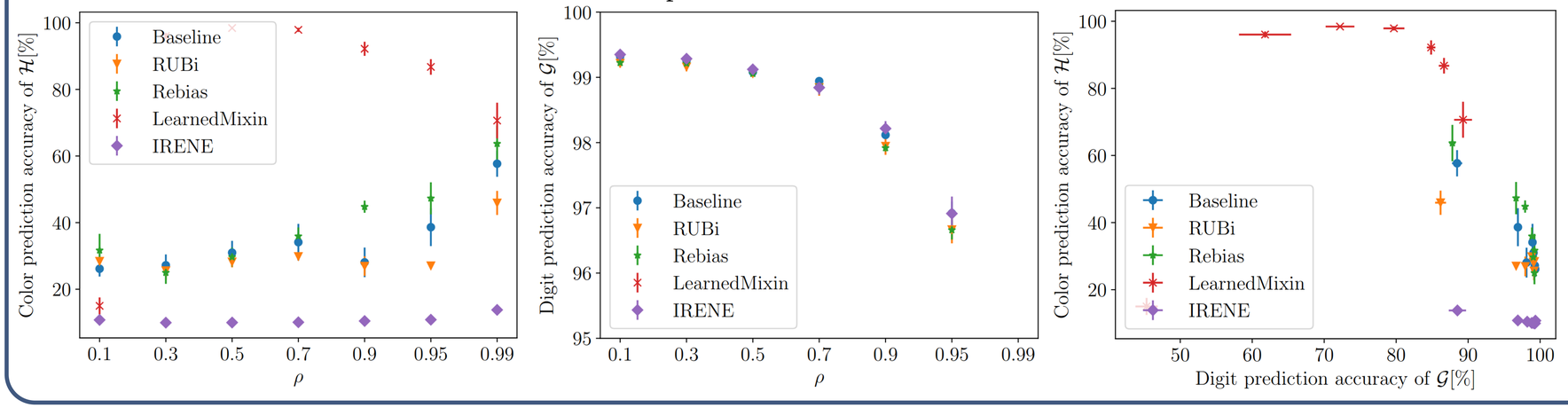
$$\mathcal{I}(v,\hat{v}) = \sum_i \sum_j p_{\sigma(v)\hat{v}}(i,j) \log \left(\frac{p_{\sigma(v)\hat{v}}(i,j)}{p_{\sigma(v)}(i) p_{\hat{v}}(j)} \right)$$

At back-propagation, the task-related loss is back-propagated through the task-related head and the backbone.
The information removal loss is back-propagated through the IR head only.
The mutual information term is back-propagated to the backbone, passing through the IR head. In the latter, no gradient will be accumulated.



Experiments

Experiments on BiasedMNIST



Experiments on CelebA

Target	Method	Prediction accuracy of $\mathcal{G}$ (trained task) [%](↑)	Gender prediction accuracy of $\mathcal{H}$ (information to remove) [%](↓)
Blond hair	Baseline	95.34±0.07	84.32±2.76
	RUBi	95.29±0.14	88.55±1.22
	Rebias	**95.59±0.11**	88.50±3.78
	LearnedMixin	90.01±2.66	74.09±2.66
	IRENE ($\gamma = 0.1$)	95.37±0.10	55.47±8.18
	IRENE ($\gamma = 0.5$)	95.28±0.09	53.64±10.69
	IRENE ($\gamma = 1$)	95.24±0.29	**53.58±10.71**
Heavy makeup	Baseline	**90.58±0.14**	92.89±0.36
	RUBi	90.40±0.08	95.17±1.11
	Rebias	90.28±0.34	93.78±2.55
	LearnedMixin	84.88±3.28	68.09±10.55
	IRENE ($\gamma = 0.1$)	90.32±0.97	65.13±11.08
	IRENE ($\gamma = 0.5$)	85.66±2.80	56.45±9.46
	IRENE ($\gamma = 1$)	83.31±3.41	**51.98±9.56**
Eyeglasses	Baseline	99.67±0.02	69.51±4.33
	RUBi	99.64±0.01	59.21±5.22
	Rebias	99.65±0.01	76.61±8.21
	LearnedMixin	93.54±2.94	61.35± 3.67
	IRENE ($\gamma = 0.1$)	99.68±0.01	64.08±1.08
	IRENE ($\gamma = 0.5$)	**99.69±0.02**	61.57±6.95
	IRENE ($\gamma = 1$)	99.68±0.01	**54.66±12.58**