REPAIR: Removing Representation Bias by Dataset Resampling

Yi Li and Nuno Vasconcelos

Statistical Visual Computing Lab, Department of Electrical and Computer Engineering, UC San Diego

Overview

- We present REPAIR, a resampling approach to minimizing representation bias of datasets
- Sensitivity analysis on video action recognition reveals some algorithms are more prone to biases than others
- Neural network models trained on de-biased datasets are shown to generalize better

Introduction

- Video action classification can often be solved with static frames with no temporal information (Fig. 1)
- Representation bias [3]: "Preference" of dataset towards different types of features
  - High bias — Feature representation informative for classification
  - Problematic if feature of high bias is not supposed to be sufficient (e.g. static features for video classification)
  - Shortcuts (visual cues) might be exploited by discriminative models (e.g. background objects, environment)
- Neural nets may overfit to bias specific to one dataset, producing unfair decisions and failing to generalize

Figure 1: Video snapshots of Kinetics [1], easily giving away the true action classes. No temporal reasoning needed here.

- Our goals:
  - Develop an algorithm (REPAIR) to reduce static bias of datasets
  - Re-evaluate action recognition models in the absence of bias
  - Improve generalization of networks using REPAIRed training set

References

[1] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[2] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, et al. HMDB: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

[3] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards action recognition without representation bias. In *ECCV*, pages 513–528, 2018.

[4] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

REPresentAtion bIas Removal (REPAIR)

Formulation Bias of dataset $\mathcal{D} = \{(X, Y)\}$ towards representation ϕ is

$$\mathcal{B}(\mathcal{D}, \phi) = 1 - \frac{\mathcal{R}^*(\mathcal{D}, \phi)}{\mathcal{H}(Y)} \tag{1}$$

with linear classification risk and label entropy

$$\mathcal{R}^*(\mathcal{D}, \phi) = \min_\theta \mathbb{E}_{X,Y}[-\log p_\theta(Y \mid \phi(X))] \approx \min_\theta -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \log p_\theta(y_i \mid \phi(x_i))$$

$$\mathcal{H}(Y) = \mathbb{E}_{X,Y}[-\log p(Y)] \approx -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \log p_{y_i}$$

- Low risk $\mathcal{R}^*(\mathcal{D}, \phi) \downarrow 0 \implies \phi$ informative for solving dataset $\mathcal{D}$, hence higher bias
- High risk $\mathcal{R}^*(\mathcal{D}, \phi) \uparrow \mathcal{H}(Y) \implies \phi$ provides little information about label Y, hence lower bias
- Goal: Obtain a new dataset $\mathcal{D}'$ derived from $\mathcal{D}$ with reduced bias

Dataset resampling Weight each example $(x_i, y_i) \in \mathcal{D}$ by its probability w_i of being selected

- Minimize reweighted bias $\mathcal{B}(\mathcal{D}'_w, \phi) = 1 - \frac{\mathcal{R}^*(\mathcal{D}'_w, \phi)}{\mathcal{H}(Y'_w)}$ with

$$\mathcal{R}^*(\mathcal{D}'_w, \phi) = \min_\theta -\sum_{i=1}^{|\mathcal{D}|} \frac{w_i}{\sum_i w_i} \log p_\theta(y_i \mid \phi(x_i))$$

$$\mathcal{H}(Y'_w) = -\sum_{i=1}^{|\mathcal{D}|} \frac{w_i}{\sum_i w_i} \log p'_{y_i}$$

$$p'_y = -\frac{\sum_{i: y_i = y} w_i}{\sum_i w_i}$$

- Leads to solving minimax problem with adversarial training

$$\min_w \max_\theta \mathcal{V}(w, \theta) = 1 - \frac{\sum_i w_i \log p_\theta(y_i \mid \phi(x_i))}{\sum_i w_i \log p'_{y_i}} \tag{2}$$

  - Classifier θ tries to classify examples in feature space ϕ
  - Weights w tries to select difficult set of examples

Colored MNIST

Experiment setup Introduce color bias to MNIST dataset by digit-dependent coloring (Fig. 2)

$$X^{\text{color}}_{i,j,c} = S_c \cdot X_{i,j}, \qquad i,j \in \{0, \ldots, 27\}, \quad c \in \{0,1,2\} \tag{3}$$

- Augment original grayscale images $X_{i,j} \in [0,1]$ with RGB color $S = (S_0, S_1, S_2) = \phi(X^{\text{color}})$
- New dataset is biased if color S dependent on class label Y (e.g. Gaussian with different mean per-class)

Resampling the Digits

Resampling strategies Selecting examples based on w_i

- `threshold` — Keep (x_i, y_i) with $w_i \geq t$
- `rank` — Keep ratio of (x_i, y_i) with greatest w_i
- `cls-rank` — Keep (x_i, y_i) with greatest w_i each class
- `sample` — Keep (x_i, y_i) with probability w_i
- `uniform` (baseline) — Pick 50% of (x_i, y_i) at random

Figure 2: Colored MNIST examples before & after resampling.

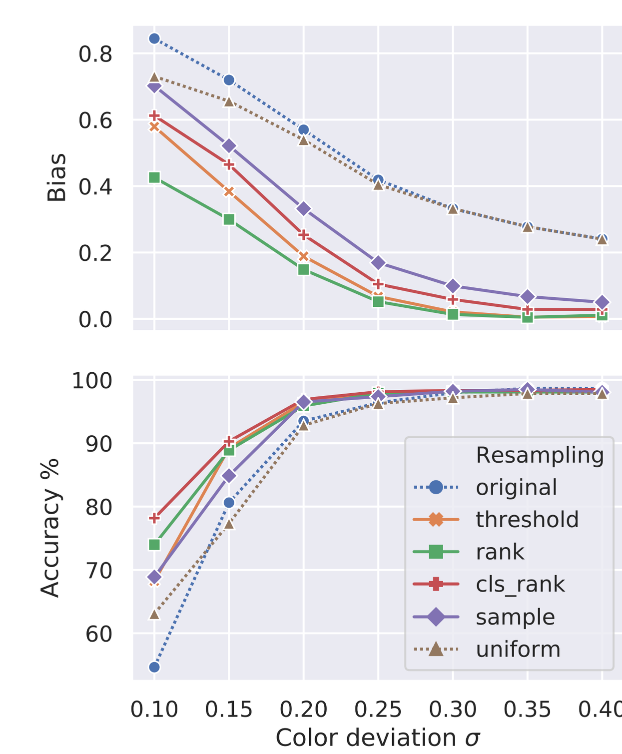

Figure 3: Bias and generalization accuracy after resampling.

Video Action Recognition

Static bias ϕ — ImageNet pre-trained CNN features

- High bias $\implies$ More static cue
- Minimize bias $\implies$ Emphasis on temporal modeling

REPAIRed dataset UCF101 [4], HMDB51 [2], Kinetics [1]

- Videos that contained too many visual cues, e.g. Billiards, are discarded (Fig. 4)
- Remaining examples are difficult for spatial CNN
- Some video CNN models rely heavily on static bias, some less (Fig. 5)

(a) UCF101. (b) HMDB51. (c) Kinetics.

Figure 4: Videos with highest/lowest resampling weights w_i.

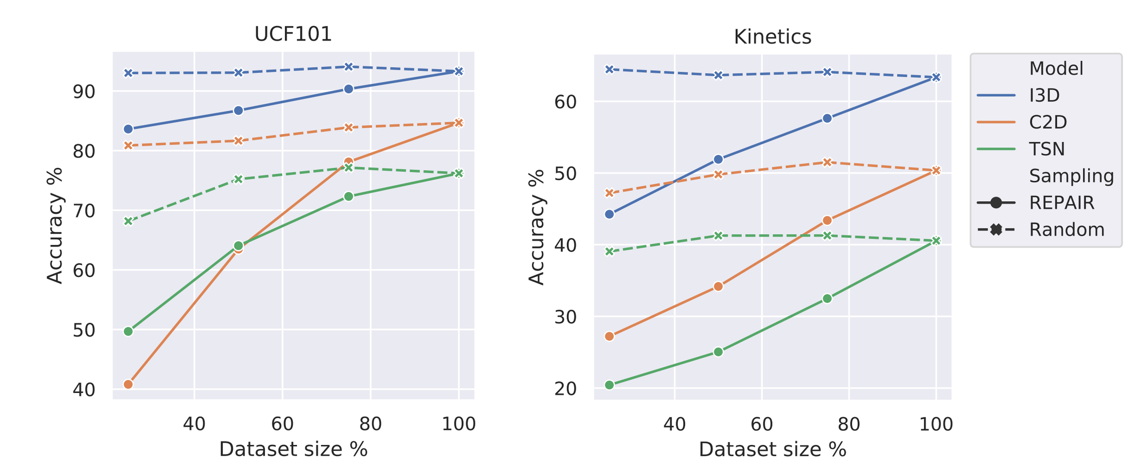

Figure 5: Algorithm performances on resampled dataset.

Generalization Training models on REPAIRed dataset

- Overfitting to the bias may hurt generalization
- 3D CNN models trained on REPAIRed Kinetics dataset generalize better to same classes in HMDB51 (Table 1)

Remove ratio Static bias	0 (orig.) 0.585	0.25 0.499	0.5 0.400	0.75 0.297
`sword`	12.43%	15.52%	16.99%	**22.03%**
`hug`	14.97%	16.26%	**17.37%**	17.11%
`somersault`	23.06%	23.97%	26.67%	**29.26%**
`laugh`	37.15%	**56.09%**	49.51%	50.42%
`clap`	**53.47%**	52.79%	52.31%	45.92%
`shake hands`	57.80%	60.31%	60.41%	**61.40%**
`kiss`	80.59%	**80.87%**	79.20%	78.96%
`smoke`	**83.31%**	80.87%	82.35%	83.29%
`pushup`	**90.70%**	88.12%	90.16%	87.26%
`situp`	90.39%	91.67%	88.09%	**92.14%**
`ride bike`	93.89%	**94.94%**	93.60%	91.02%
`pullup`	100.00%	100.00%	100.00%	100.00%
Average	61.48%	**63.45%**	63.06%	63.24%

Table 1: Cross-dataset generalization from REPAIRed Kinetics to HMDB51.