Scalable Ensemble Diversification for OOD Generalization and Detection


Alexander Rubinstein


Luca Scimeca


Damien Teney


Seong Joon Oh

STAI
imprs-is
Tübingen AI Center
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
Mila
idiap RESEARCH INSTITUTE

Shortcut biases are reason for failures in computer vision

Background bias in image classification:

Predicted as Cow

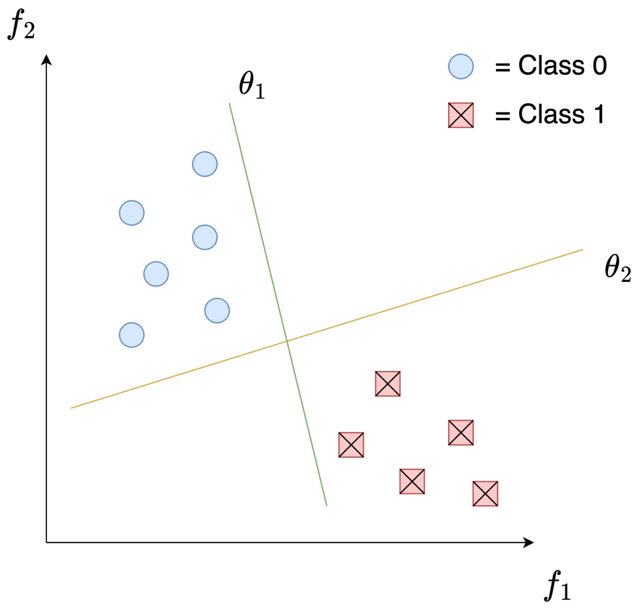
Predicted as Dolphin

Shortcut biases are reason for failures in natural language processing

Parametric answer bias in question answering for retrieval augmented language models:

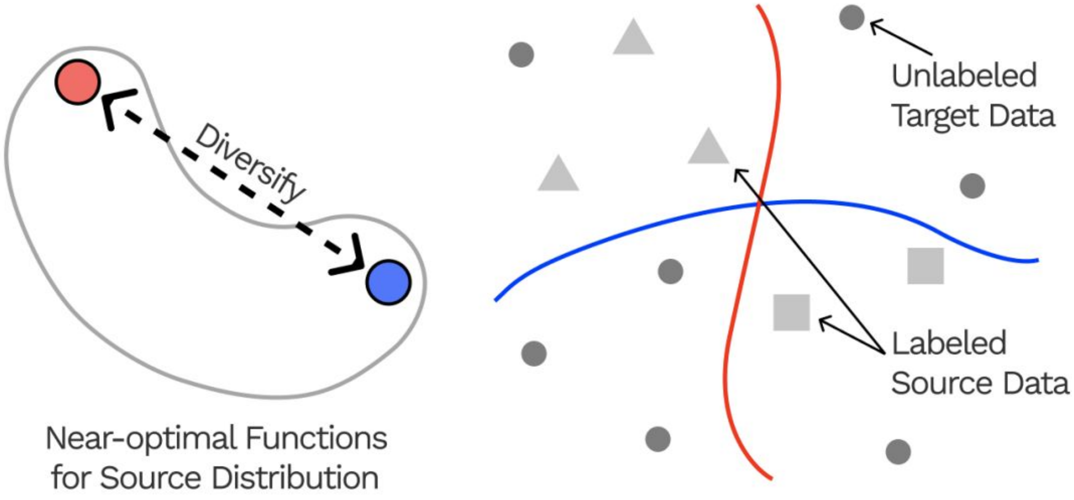
Question: Who was the main performer at this year's halftime show?
Document: CBS broadcast Super Bowl 50 in the U.S., and charged an average of $5 million for a 30-second commercial during the game. The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively. It was the third-most watched U.S. broadcast ever.
Ground-truth answer: Coldplay
Incorrect parametric answer: Beyoncé

Intuition for mitigating shortcut biases by multiple hypotheses



Underspecified dataset = several features can be used to predict ground truth

Current diversification approaches were designed for small-scale datasets



$$\mathcal{L} = \mathcal{L}_{\text{agree}}(\mathcal{D}_{\text{ID}}) + \mathcal{L}_{\text{disagree}}(\mathcal{D}_{\text{OOD}})$$

Diversify and Disambiguate: Learning From Underspecified Data (ICLR 2023)

On ImageNet scale they do not work

$$\mathcal{L} = \mathcal{L}_{\text{agree}}(\mathcal{D}_{\text{ID}}) + \mathcal{L}_{\text{disagree}}(\mathcal{D}_{\text{OOD}})$$

Method	$\mathcal{D}_{\text{OOD}}$	IN-val	IN-A	IN-R
Deep ensemble	-	**85.4**	**39.9**	46.3
+Diverse HPs	-	**85.4**	**39.9**	**46.5**
A2D	IN-A	85.1	37.8	45.2
A2D	IN-R	85.1	37.8	45.2
Div	IN-A	85.1	37.8	45.2
Div	IN-R	85.1	35.7	41.8

On which samples models tend to disagree?

y: corkscrew
first_pred: corkscrew
second_pred: wine bottle
index: 70609

Multilabel

y: assault rifle
first_pred: rifle
second_pred: assault rifle
index: 61206

Subclass relationship

y: Indian elephant
first_pred: tusker
second_pred: Indian elephant
index: 51162

Easy to confuse

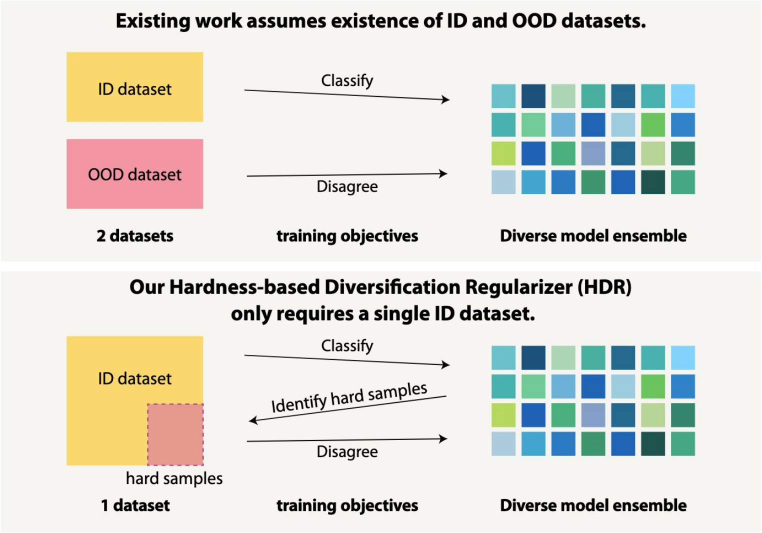
How to find samples for disagreement within in-distribution (ID) data?

Idea: identify hard training samples with high cross-entropy (CE) and disagree on them

Do it via adaptive reweighting:

$$\mathcal{L} = \mathcal{L}_{\text{agree}}(\mathcal{D}_{\text{ID}}) + \alpha \cdot \mathcal{L}_{\text{disagree}}(\mathcal{D}_{\text{ID}})$$

- In the beginning of training α is almost 0
- On the later epochs, for each sample α is proportional to CE on this sample



Details on loss with adaptive weights. Hardness-based diversification regularizer (HDR).

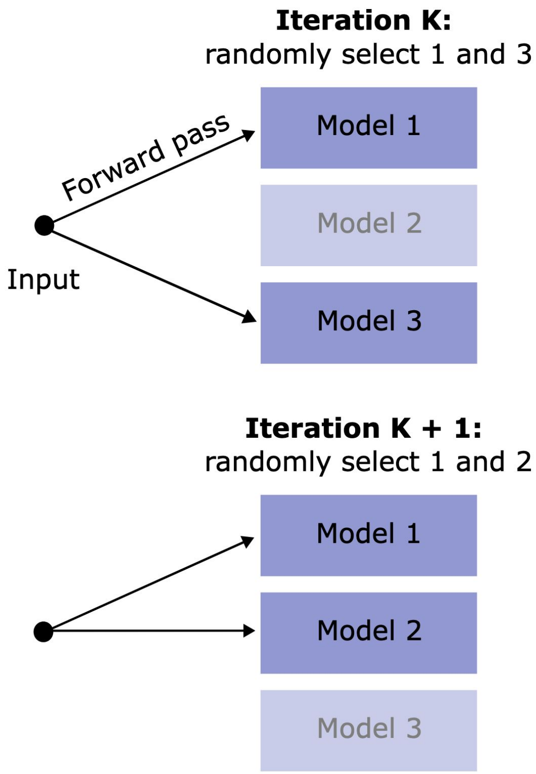
$$\alpha_n := \frac{\mathrm{CE}\left(\frac{1}{M}\sum_m f^m\left(x_n\right), y_n\right)}{\left(\frac{1}{|B|}\sum_{b\in B}\mathrm{CE}\left(\frac{1}{M}\sum_m f^m\left(x_b\right), y_b\right)\right)^2}$$

$$\mathcal{G}\left(p^m(x), p^l(x)\right) = -\log\left(p^m_{\hat{y}}(x)\cdot\left(1-p^l_{\hat{y}}(x)\right) + p^l_{\hat{y}}(x)\cdot\left(1-p^m_{\hat{y}}(x)\right)\right)$$

$$\mathcal{L}_{\text{main}} = \frac{1}{MN}\sum_n^N\sum_m^M -\log p^m_{y_n}\left(x_n;\theta\right)$$

$$\mathcal{L}_{\text{HDR}} := \mathcal{L}_{\text{main}} + \frac{\lambda}{NM(M-1)}\sum_n\sum_{m<l}\operatorname{stopgrad}\left(\alpha_n\right)\cdot\mathcal{G}\left(p^m\left(x_n\right), p^l\left(x_n\right)\right)$$

Stochastic sum allows to train ensembles of any size



Results in OOD generalization

Method	#Models	Val	IN-A	IN-R
Deep ensemble	5	**85.4**	39.9	46.3
+Diverse HPs	5	**85.4**	39.9	46.5
DivDis	5	85.1	36.3	41.8
A2D	5	85.1	37.8	45.2
HDR (Ours)	5	85.3	**43.0**	**48.7**
Deep ensemble	50	**85.5**	38.8	45.8
+Diverse HPs	50	**85.5**	42.5	48.5
HDR (Ours)	50	83.6	**50.6**	**53.8**

Novel way to measure epistemic uncertainty

Idea: measure diversity of outputs as number of uniquely predicted classes instead of ensemble confidence ($\bar{p}$)

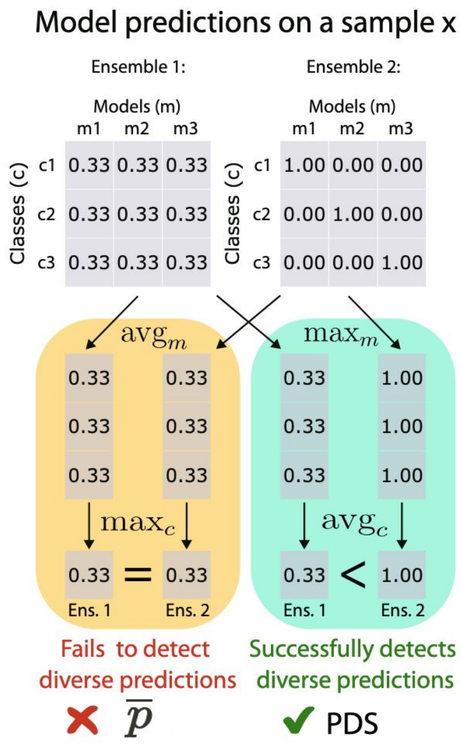
Discrete formula:

$$\hat{Y} = \{\text{argmax}_c\, p_c^m(x), m = 1 \dots M\}$$

$$\eta_{\#\text{unique}} := \frac{1}{C}\,\text{num\_unique}(\hat{Y})$$

Continuous approximation - predictive diversity score (PDS):

$$\eta_{\text{PDS}} := \frac{1}{C}\sum_c \max_m p_c^m(x)$$



Results in OOD detection

Models	η	C-1	C-5	iNaturalist	OpenImages
Single model	p	61.5	83.3	95.8	90.9
Deep Ensemble	$\overline{p}$	61.9	83.5	95.8	91.1
+Diverse HPs	$\overline{p}$	**64.2**	**86.1**	**96.9**	**92.3**
DivDis	$\overline{p}$	59.8	84.3	96.6	92.2
A2D	$\overline{p}$	59.4	83.5	96.6	91.6
HDR (Ours)	$\overline{p}$	64.1	84.5	96.0	91.5
Deep Ensemble	PDS	56.5	62.5	59.2	58.9
+Diverse HPs	PDS	64.3	84.9	92.6	88.9
DivDis	PDS	60.0	85.1	96.9	93.9
A2D	PDS	59.9	85.0	97.1	93.9
HDR (Ours)	PDS	**68.1**	**89.4**	**97.7**	**94.1**

Conclusions

- Identifying samples for disagreement within ID data + stochastic sum enables scaling of diverse ensembles to ImageNet
- Diversify ensembles by making them disagree on hard samples
- Use PDS to measure epistemic uncertainty and detect OOD samples

Appendix

Ensemble benefits from diversification

When we average outputs of multiple models error is:

$$\mathrm{Err}(\overline{f}) = \overline{\mathrm{Err}(f)} - \mathrm{Var}\, f$$

Error of averaged model

Mean Error of single model

Variance of model outputs

If we want to make $\mathrm{Err}(\overline{f})$ small

For that we need to increase $\mathrm{Var}\, f$

While keeping $\overline{\mathrm{Err}(f)}$ low

More formally

$$\overline{\boldsymbol{f}}(\boldsymbol{x}) = \mathop{\mathbb{E}}_{p(f)}[f(\boldsymbol{x})]$$

$$\mathrm{Var}_{p(\boldsymbol{f})}[\boldsymbol{f}(\boldsymbol{x})] = \sum_{i=1}^{C} \mathrm{Var}_{p(\boldsymbol{f})}\left[f^{(i)}(\boldsymbol{x})\right]$$

$$\mathrm{B}(\boldsymbol{f}(\boldsymbol{x}), y) = \mathop{\mathbb{E}}_{p(x)}\left[\sum_{i=1}^{C}\left(f^{(i)}(x) - y^{(i)}\right)^2\right]$$

$$\mathop{\mathbb{E}}_{p(\boldsymbol{f})}[\mathrm{B}(\boldsymbol{f}(\boldsymbol{x}), y)] - \mathrm{B}(\overline{\boldsymbol{f}}(\boldsymbol{x}), y) = \mathop{\mathbb{E}}_{p(x)}[\mathop{\mathrm{Var}}_{p(\boldsymbol{f})}[\boldsymbol{f}(\boldsymbol{x})]]$$

Stochastic sum is an unbiased estimator

$$\mathcal{L} = \mathcal{L}_{\text{agree}} + L_{\text{disagree}} = \frac{1}{|M|} \sum_{m \in M} \mathcal{L}(m) + \frac{1}{|P_M|} \sum_{p \in P_M} \mathcal{G}(p).$$

$$\overline{\nabla \mathcal{L}_{agree}} = \frac{1}{|I|} \sum_{m \in I} \nabla \mathcal{L}(m)$$

$$\mathbb{E}_{m \in M} \left[\overline{\nabla \mathcal{L}_{agree}} \right] = \frac{1}{|I|} \sum_{m \in I} \mathbb{E}_{m \in M} [\nabla \mathcal{L}(m)] = \frac{1}{|I|} \cdot |I| \frac{1}{M} \sum_{m \in M} \nabla \mathcal{L}(m) = \nabla \left[\frac{1}{M} \sum_{m \in M} \mathcal{L}(m) \right] = \nabla \mathcal{L}_{\text{agree}}$$

$$\overline{\nabla \mathcal{L}_{\text{disagree}}} = \frac{1}{|I|} \sum_{p \in I} \nabla \mathcal{G}(p)$$

$$\mathbb{E}_{m \in M} \left[\overline{\nabla \mathcal{L}_{\text{disagree}}} \right] = \frac{1}{|\eta|} \sum_{p \in P_I} \mathbb{E}_{m \in M} [\nabla \mathcal{G}(p)] = \frac{1}{|I|} \cdot |I| \cdot \frac{1}{|P_M|} \sum_{p \in P_M} \nabla \mathcal{G}(p) = \nabla \mathcal{L}_{\text{disagree}}$$