Enhancing Tail Performance In Extreme Classifiers by Label Variance Reduction

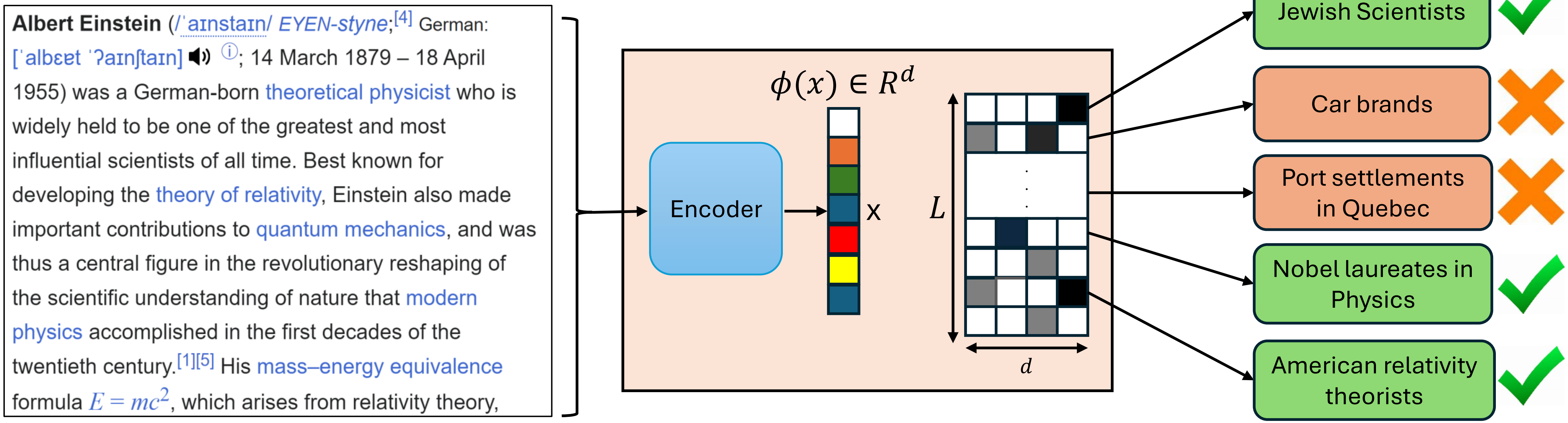
Anirudh Buvanesh, Rahul Chand, Jatin Prakash, Bhawna Paliwal, Mudit Dhawan, Neelabh Madan, Deepesh Hada, Vidit Jain, Sonu Mehta, Yashoteja Prabhu, Manish Gupta, Ramachandran Ramjee, Manik Varma

Anirudh Buvanesh
Research Fellow, Microsoft Research India

Extreme Classification

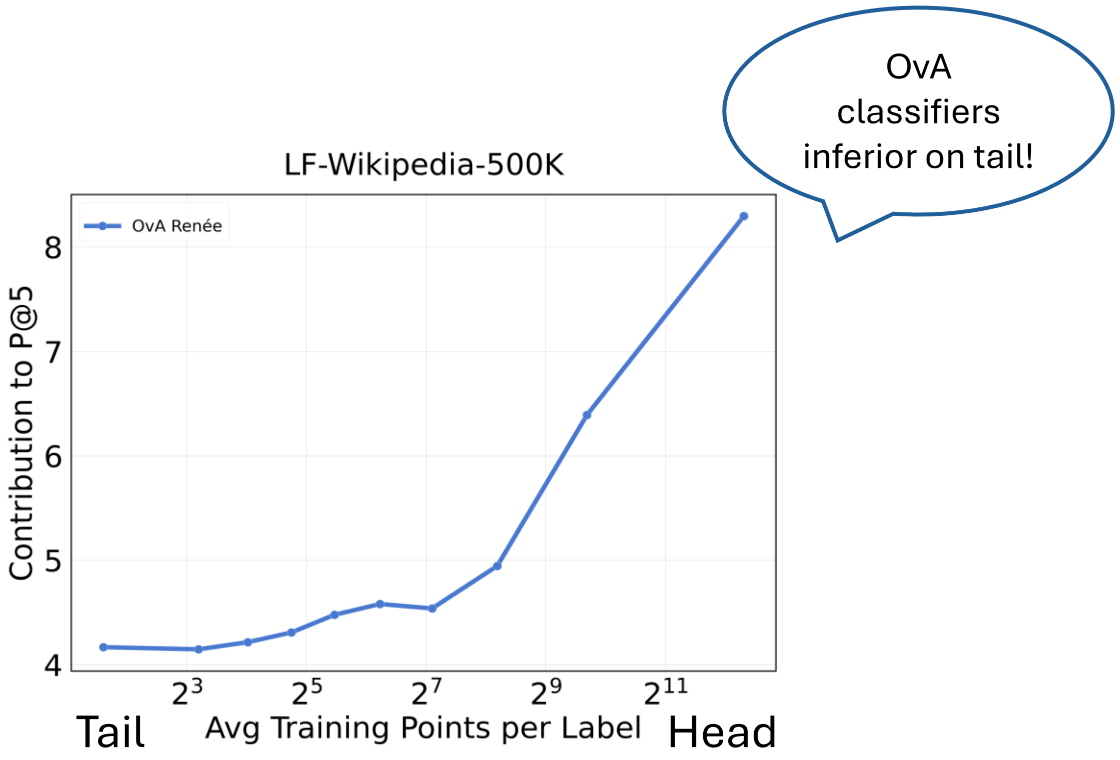
- Goal: Map a data point to the most relevant subset of labels.

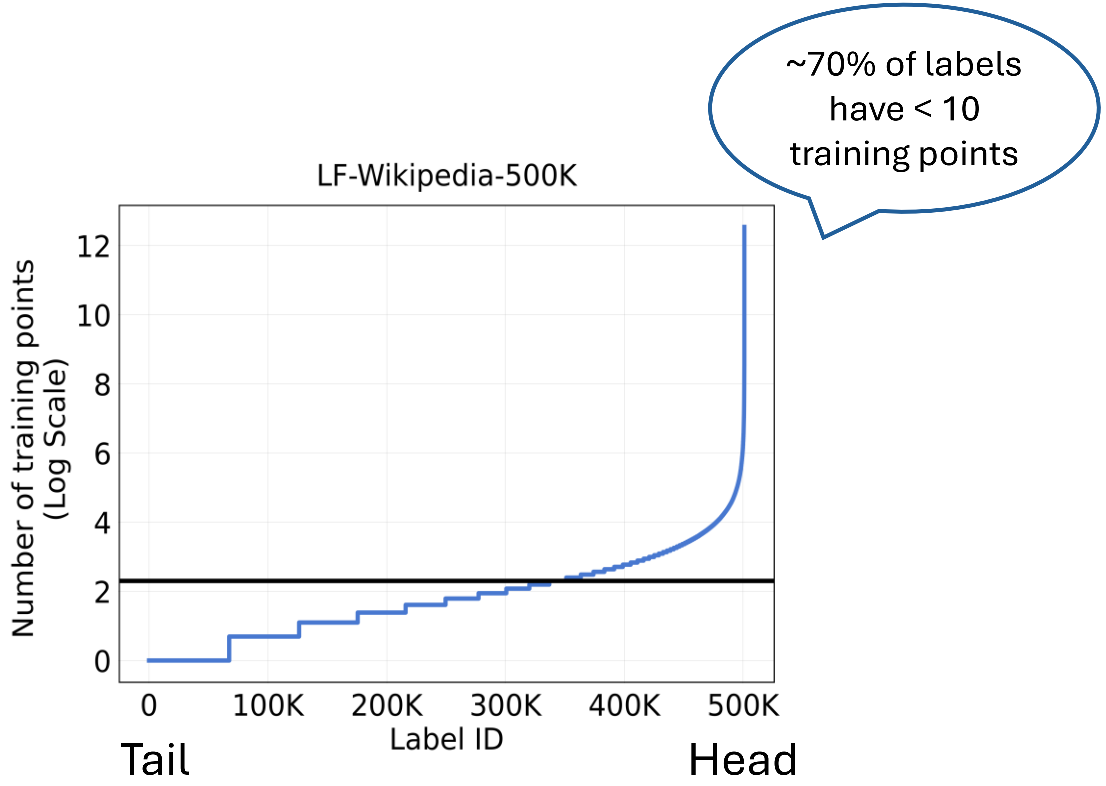
Wikipedia article on Albert Einstein



Tail Labels

- OvA classifiers overfit on tail labels
- Tail Label constitute a significant proportion of labels





Mitigating inferior tail performance

- Regularizing OvA Classifiers
- Loss Reweighting
- Augmenting tail labels

Label Variance

- Measures imprecision in dataset
- Difference in annotator’s judgments
- Fluctuation in user interests

Effect of Label Variance

- *Theorem 1:*
  - Generalization performance of OvA Classifiers $\propto \frac{1}{Label\ Variance}$
  - Precise relevance estimates ($\mathbb{P}(\text{Label is relevant} \mid \text{Query})$), lowers Label variance
- *Lemma 1:*
  - Upper bound on label variance is $\propto \frac{1}{Number\ of\ samples\ for\ label}$
  - Tail Labels have high label variance

Goal: Reducing label variance using a teacher model

Reducing Label Variance

- Siamese networks perform better on tail labels
- Relevance scores from Siamese networks are not well calibrated
- *Theorem 3:* Training Siamese teacher ($\mathcal{E}_\theta$) with logistic loss gives well calibrated relevance estimates

$$\mathcal{L}_{\text{logistic}} = \min_\theta \sum_{i \in \text{data points}} \sum_{\substack{a \in \text{postives} \\ b \in \text{hard negs}}} \log(1 + e^{\mathcal{E}_\theta(i)^\top \mathcal{E}_\theta(b) - \mathcal{E}_\theta(i)^\top \mathcal{E}_\theta(a)})$$

- *Theorem 2:* Training OvA classifiers with ground truth and teacher relevance estimates is more optimal that using either one alone.

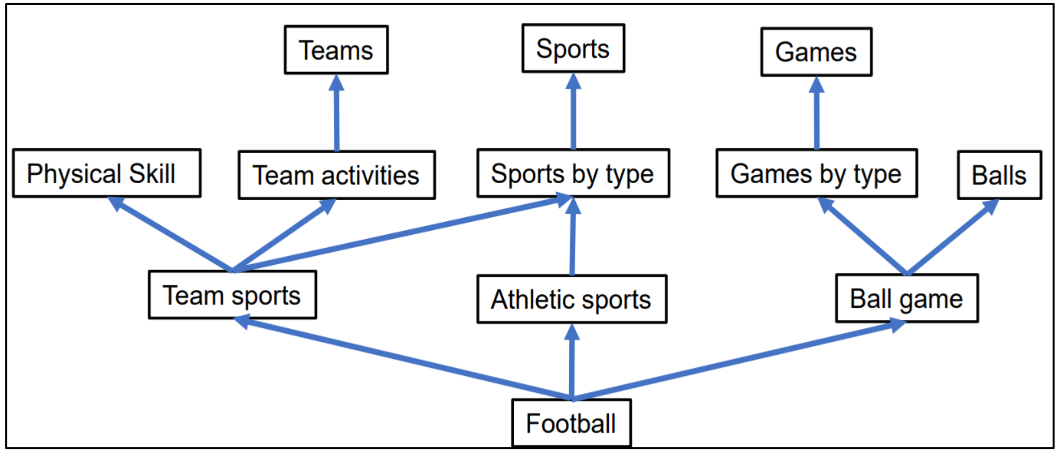
$$\mathcal{L}_{\text{lever}} = \lambda\, \mathcal{L}_{\text{bce}}(y_l, \mathbf{w}_l^\top \mathbf{x}) + (1-\lambda)\, \mathcal{L}_{\text{bce}}(\hat{p}_l, \mathbf{w}_l^\top \mathbf{x});\;\; \hat{p}_l = \mathcal{E}_\theta(\mathbf{x})^\top \mathcal{E}_\theta(l)$$

New Datasets

- Queries and Labels have lesser semantic overlap
- Longer tail
- Resemble real world applications

Query Completion Task (AOL Dataset)

Intent Generalization Task (Wiki-Hierarchy Dataset)



Results

- LEVER can easily combine with any OvA classifier
- Improves Precision by **1.4%**, PSP by **5%**, Coverage by **6.5%**
- No inference overhead, training time is at most 2x for Renée

Model	**LF-AmazonTitles-131K**		**LF-Wikipedia-500K**		**LF-AOL-270K**	
	P@5	PSP@5	P@5	PSP@5	P@5	PSP@5
ELIAS	18.14	39.08	48.75	48.67	14.91	25.22
ELIAS + LEVER	**20.16**	**45.43**	**50.03**	**55.03**	**15.57**	**30.43**
CascadeXML	18.18	38.81	45.10	43.29	14.82	23.19
CascadeXML + LEVER	**20.63**	**46.95**	**46.44**	**50.99**	**14.99**	**27.59**
Renée	**22.04**	**50.33**	51.68	55.68	15.85	32.19
Renée + LEVER	21.92	50.31	**51.98**	**60.29**	**17.07**	**45.13**

Thank You!

- Paper: https://openreview.net/pdf?id=6ARlSgun7J
- Code: https://github.com/anirudhb11/LEVER
- Reach out: anirudhb1102@gmail.com; yprabhu@microsoft.com