Outlier-Aware Training for Improving Group Accuracy Disparities

Li-Kuang Chen Canasai Kruengkrai Junichi Yamagishi

國立清華大學
NATIONAL TSING HUA UNIVERSITY

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Background

Classifiers that use standard training can achieve high average accuracy but perform poorly on certain minority groups.

- Standard training (ERM) models often spuriously correlate attributes, such as the existence of negation words in a sentence, to frequently-co-occurring labels.

Background

Classifiers that use standard training can achieve high average accuracy but low accuracy on certain minority groups.

- We can partition a dataset into “groups”, where each group is a set of examples with a combination of a spurious *attribute* and a class *label.*

Background

Classifiers that use standard training can achieve high average accuracy but low accuracy on certain minority groups.

- We can partition a dataset into “groups”, where each group is a set of examples with a combination of a spurious *attribute* and a class *label.*

FEVER dataset (Thorne et al. 2018)

Input: claim + evidence ⇒ **Label:** $\left\{ \begin{array}{c} \text{Supports} \\ \text{Refutes} \\ \text{NotEnoughInfo} \end{array} \right\}$

Background

Classifiers that use standard training can achieve high average accuracy but low accuracy on certain minority groups.

- We can partition a dataset into “groups”, where each group is a set of examples with a combination of a spurious *attribute* and a class *label.*

FEVER dataset (Thorne et al. 2018)

Input: claim + evidence ⇒ **Label:** { Supports / Refutes / NotEnoughInfo }

Attribute: { with negation word / without negation word } in claim

Background

Classifiers that use standard training can achieve high average accuracy but low accuracy on certain minority groups.

- We can partition a dataset into “groups”, where each group is a set of examples with a combination of a spurious *attribute* and a class *label.*

FEVER dataset (Thorne et al. 2018)

Input: claim + evidence ⇒ **Label:** {Supports, Refutes, NotEnoughInfo}

Attribute: {with negation word, without negation word} in claim

6 groups

Background

Classifiers that use standard training can achieve high average accuracy but low accuracy on certain minority groups.

- We can partition a dataset into "groups", where each group is a set of examples with a combination of a spurious *attribute* and a class *label.*

FEVER dataset (Thorne et al. 2018)

Input: claim + evidence ⇒ **Label:** { Supports / Refutes / NotEnoughInfo }

claim: *"Luis Fonsi does **not** go by his given name on stage."*
evidence:
"Luis Alfonso Rodríguez López-Cepero, more commonly known by his stage name Luis Fonsi, (born April 15, 1978) is a Puerto Rican singer, songwriter and actor."

Attribute: { with negation word / without negation word } in claim

Background

Classifiers that use standard training can achieve high average accuracy but low accuracy on certain minority groups.

- Models perform poorly on groups ("worst groups") where the spurious correlation does not hold.

FEVER dataset (Thorne et al. 2018)

Input: claim + evidence ⇒ **Label:** { Supports / Refutes / NotEnoughInfo }

87.8% average accuracy

48.6% acc. on *worst group* (**Supports** with **negation**)

claim: *"Luis Fonsi does not go by his given name on stage."*

evidence:

"Luis Alfonso Rodríguez López-Cepero, more commonly known by his stage name Luis Fonsi, (born April 15, 1978) is a Puerto Rican singer, songwriter and actor."

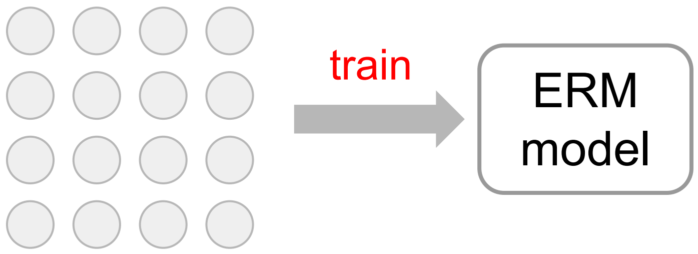
Attribute: { with negation word / without negation word } in claim

Background: Related Work

- Methods that use group and attribute information during training can effectively improve worst group accuracy, but are expensive to annotate.
- We focus on improving one of the methods that do not require group information during training, called Just Train Twice (Liu et al., 2021), or JTT.

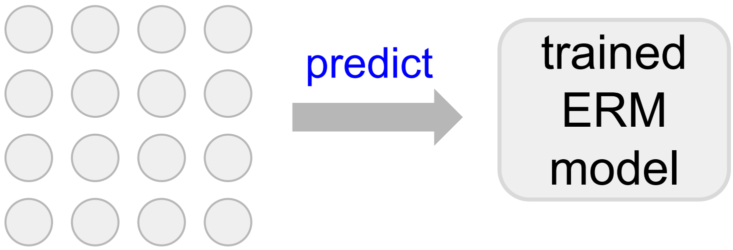
Background: JTT (Liu et al. 2021)

Training set examples



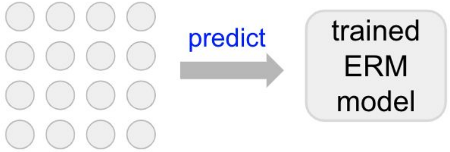
Background: JTT

Training set examples

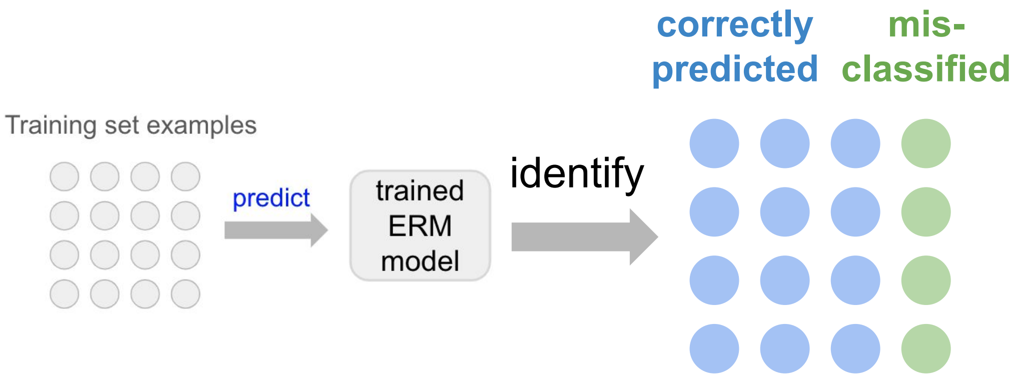


Background: JTT

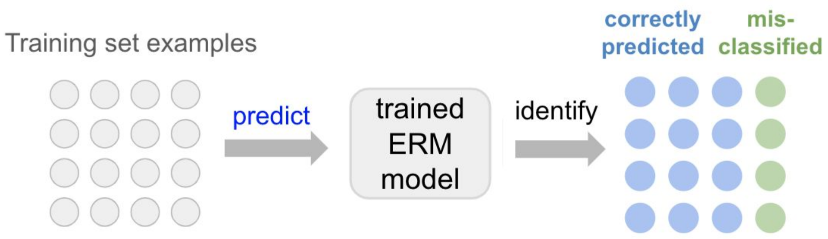
Training set examples



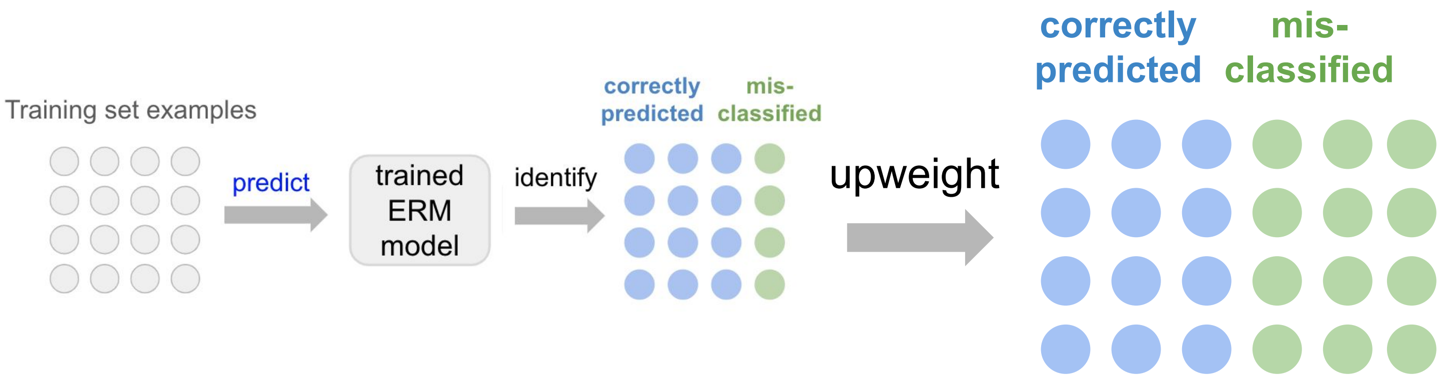
Background: JTT



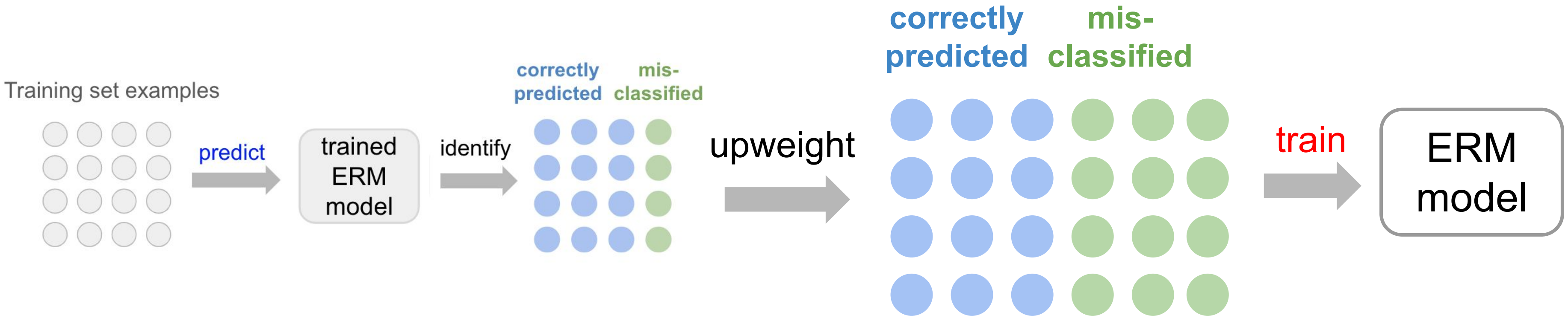
Background: JTT



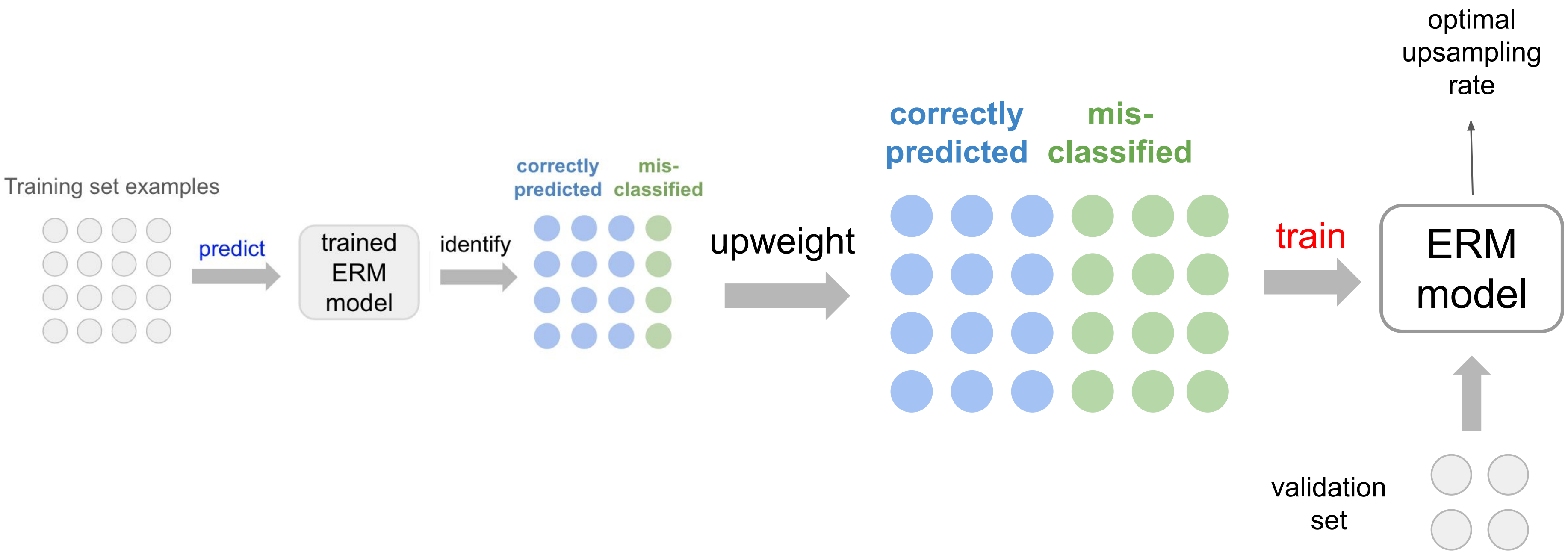
Background: JTT



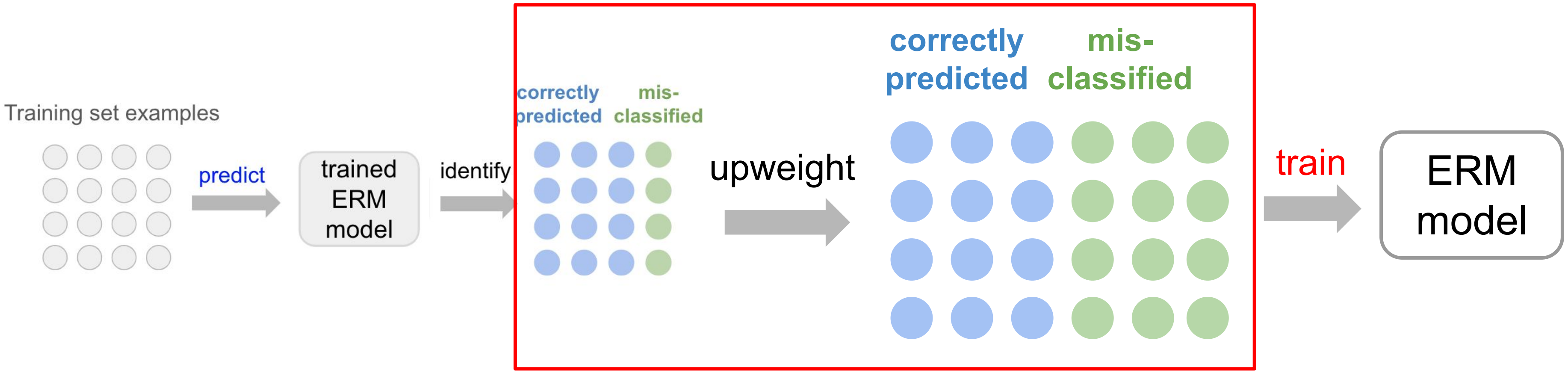
Background: JTT



Background: JTT

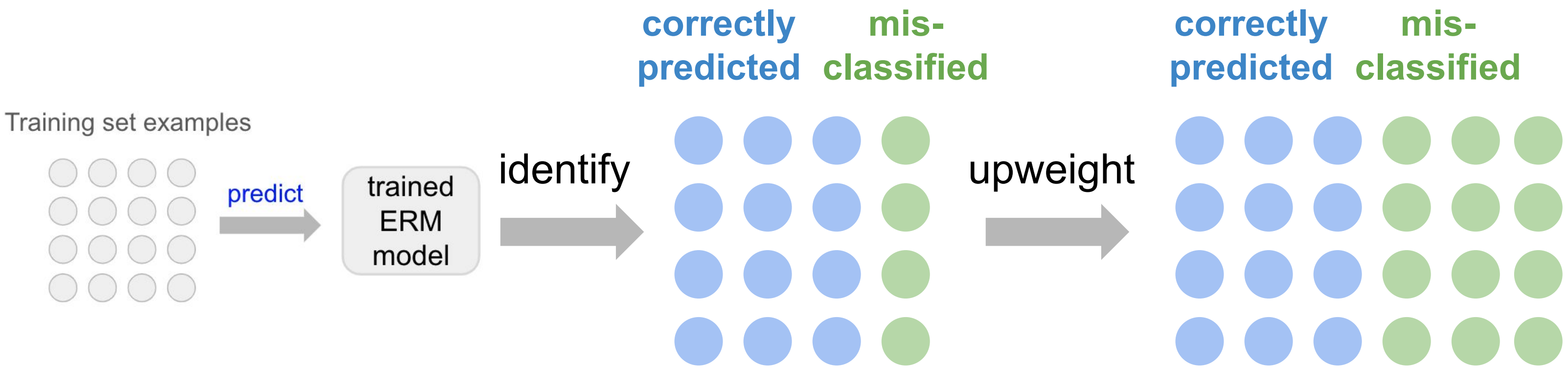


Proposed Method: JTT-m



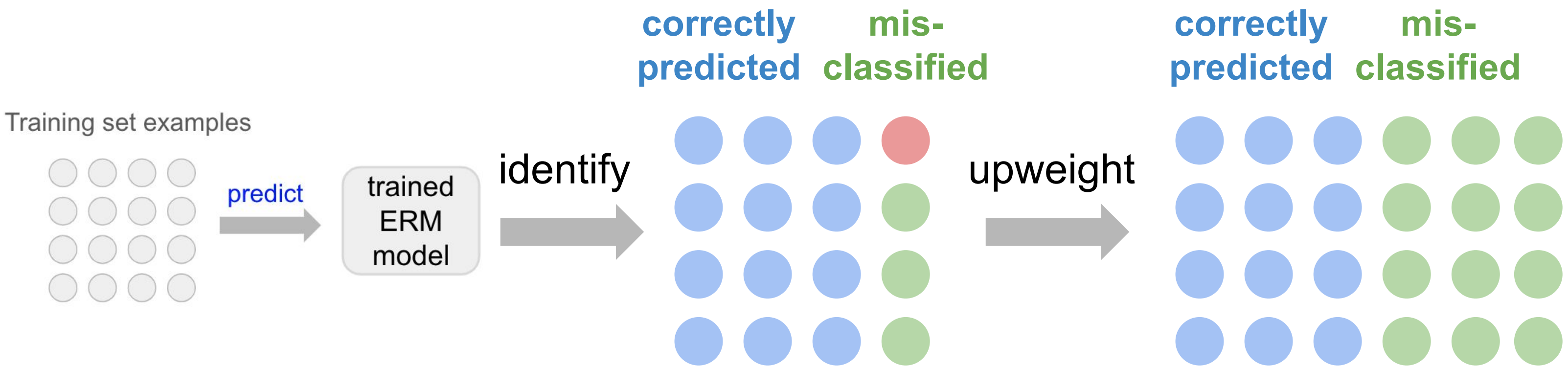
Proposed Method: JTT-m

The upweighted error set may contain untrainable or out-of-distribution (OOD) examples



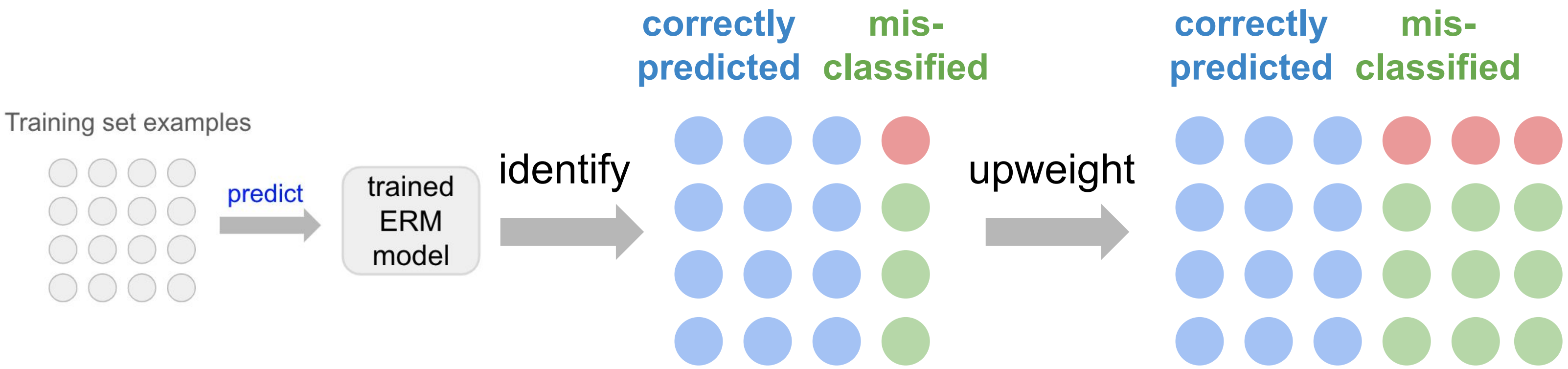
Proposed Method: JTT-m

The upweighted error set may contain untrainable or out-of-distribution (OOD) examples



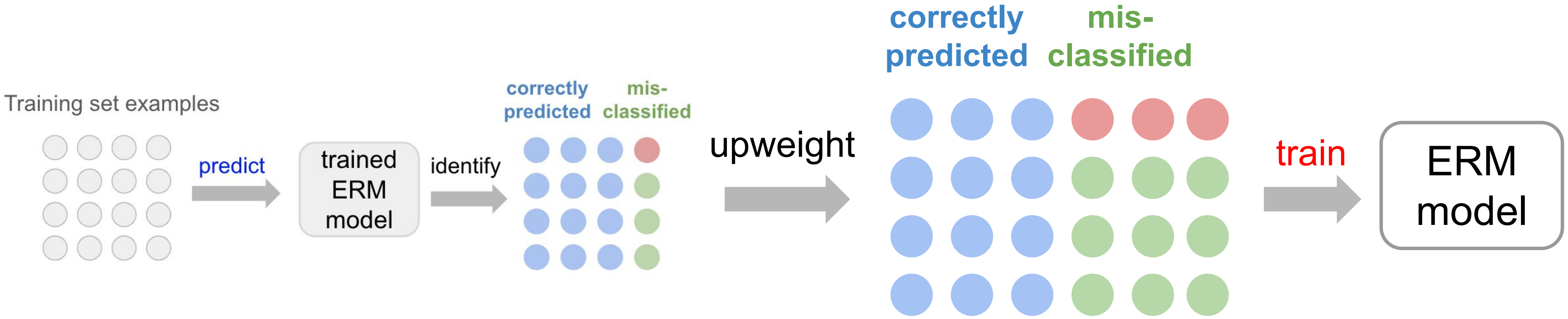
Proposed Method: JTT-m

The upweighted error set may contain untrainable or out-of-distribution (OOD) examples



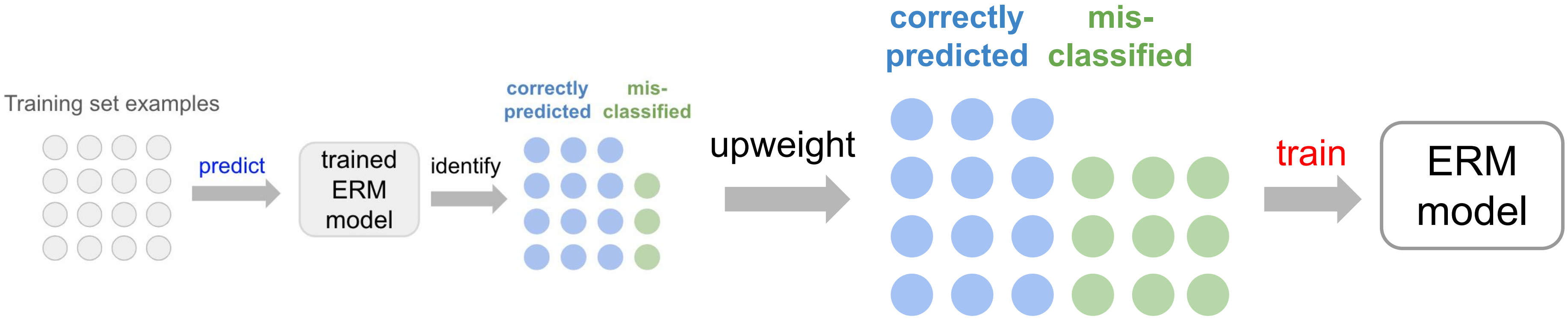
Proposed Method: JTT-m

When these undesirable examples gets upweighted, JTT's effectiveness might be hampered.



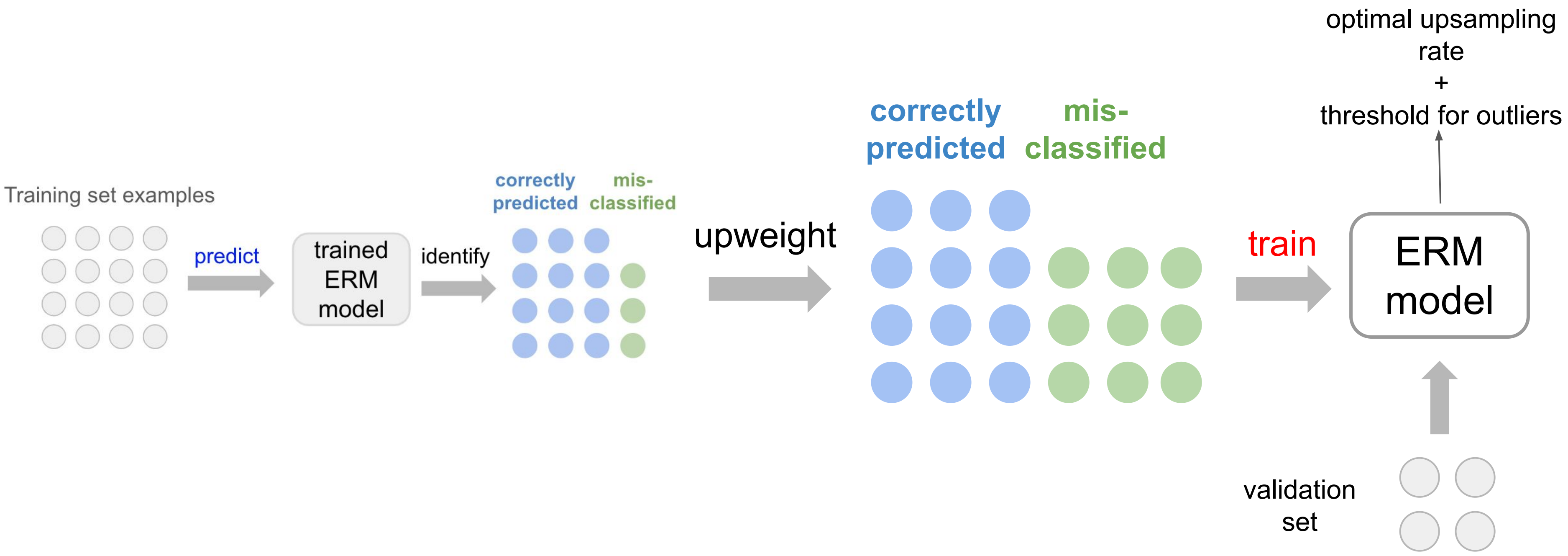
Proposed Method: JTT-m

We propose to **remove the outliers from the error set before upweighting!**



Proposed Method: JTT-m

We propose to **remove the outliers from the error set before upweighting**!



Proposed Method: JTT-m

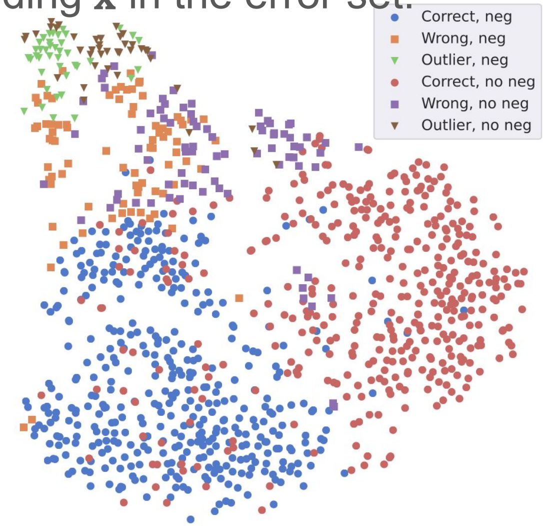
We propose to remove the outliers from the error set before upweighting!

To remove the outliers, we adopt a similar approach to Lee et al. (2018)'s method:

1. Obtain the penultimate embedding layer of the model
2. Calculate Mahalanobis distance $M(\mathbf{x})$ for each example embedding $\mathbf{x}$ in the error set:

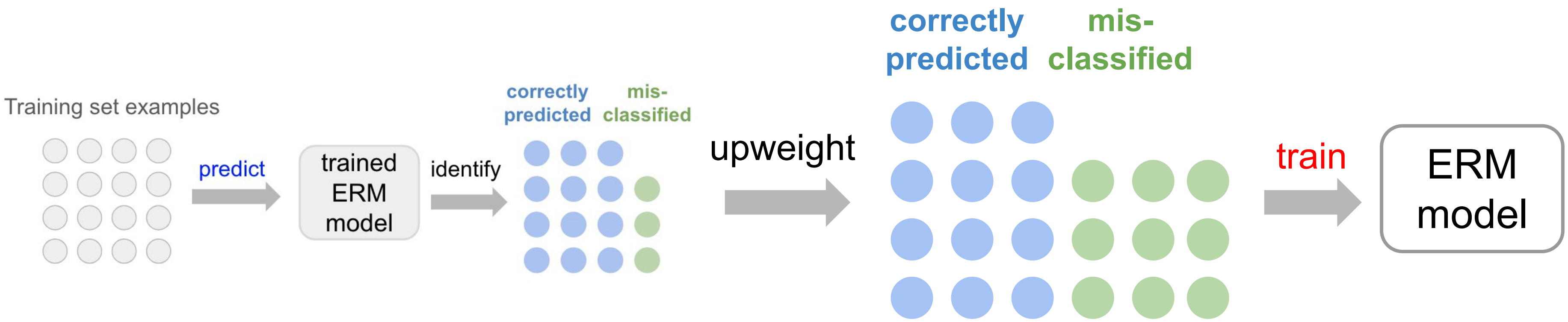
$$M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{x} - \boldsymbol{\mu}_y)}$$

3. Filter out examples whose M(x) does not meet our threshold



Proposed Method: JTT-m

We propose to **remove the outliers from the error set before upweighting!**



Experiments: Datasets

Experiments: Datasets

FEVER (Thorne et al. 2018)

Claim	Andy Roddick lost 5 Master Series between 2002 and 2010.
Evidence	Roddick was ranked in the top 10 for nine consecutive years between 2002 and 2010 , at year's end, and won five Masters Series in that period
Label	Refutes
Spurious attribute	negation word in claim

Experiments: Datasets

FEVER (Thorne et al. 2018)

Claim	Andy Roddick lost 5 Master Series between 2002 and 2010.
Evidence	Roddick was ranked in the top 10 for nine consecutive years between 2002 and 2010 , at year's end, and won five Masters Series in that period
Label	Refutes
Spurious attribute	negation word in claim

MultiNLI (Williams et al. 2017)

Hypothesis	California cannot do any better.
Premise	Clearly, California can - and must - do better.
Label	Contradiction
Spurious attribute	negation word in hypothesis

Experiments: Datasets

FEVER (Thorne et al. 2018)

Claim	Andy Roddick lost 5 Master Series between 2002 and 2010.
Evidence	Roddick was ranked in the top 10 for nine consecutive years between 2002 and 2010 , at year's end, and won five Masters Series in that period
Label	Refutes
Spurious attribute	negation word in claim

MultiNLI (Williams et al. 2017)

Hypothesis	California cannot do any better.
Premise	Clearly, California can - and must - do better.
Label	Contradiction
Spurious attribute	negation word in hypothesis

Experiments: Indicators to Track

- Worst-group accuracy
- Average accuracy

Dataset	FEVER		MultiNLI	
	Avg. (%)	Worst (%)	Avg. (%)	Worst (%)
ERM	87.8 $\pm$0.2	48.6 $\pm$0.7	84.9 $\pm$0.1	72.0 $\pm$1.0
JTT	86.8 $\pm$0.2	50.5 $\pm$3.5	83.0 $\pm$0.2	75.5 $\pm$1.5
JTT-m	87.4 $\pm$0.1*	50.2 $\pm$2.8	83.0 $\pm$0.3	77.3 $\pm$0.4*

Experiments: Indicators to Track

Dataset	FEVER		MultiNLI	
	Avg. (%)	Worst (%)	Avg. (%)	Worst (%)
ERM	$87.8_{\pm 0.2}$	$48.6_{\pm 0.7}$	$84.9_{\pm 0.1}$	$72.0_{\pm 1.0}$
JTT	$86.8_{\pm 0.2}$	$50.5_{\pm 3.5}$	$83.0_{\pm 0.2}$	$75.5_{\pm 1.5}$
JTT-m	$87.4_{\pm 0.1}$*	$50.2_{\pm 2.8}$	$83.0_{\pm 0.3}$	$77.3_{\pm 0.4}$*

Experiments: Indicators to Track

Dataset	FEVER		MultiNLI	
	Avg. (%)	Worst (%)	Avg. (%)	Worst (%)
ERM	$87.8_{\pm 0.2}$	$48.6_{\pm 0.7}$	$84.9_{\pm 0.1}$	$72.0_{\pm 1.0}$
JTT	$86.8_{\pm 0.2}$	$50.5_{\pm 3.5}$	$83.0_{\pm 0.2}$	$75.5_{\pm 1.5}$
JTT-m	$87.4_{\pm 0.1}$*	$50.2_{\pm 2.8}$	$83.0_{\pm 0.3}$	$77.3_{\pm 0.4}$*

Experiments: Indicators to Track

Dataset	FEVER		MultiNLI	
	Avg. (%)	Worst (%)	Avg. (%)	Worst (%)
ERM	$87.8_{\pm 0.2}$	$48.6_{\pm 0.7}$	$84.9_{\pm 0.1}$	$72.0_{\pm 1.0}$
JTT	$86.8_{\pm 0.2}$	$50.5_{\pm 3.5}$	$83.0_{\pm 0.2}$	$75.5_{\pm 1.5}$
JTT-m	$87.4_{\pm 0.1}$*	$50.2_{\pm 2.8}$	$83.0_{\pm 0.3}$	$77.3_{\pm 0.4}$*

* indicates statistical significance of difference between JTT and JTT-m (paired t-test, $p < 0.05$)

Experiments: Indicators to Track

Dataset	FEVER		MultiNLI	
	Avg. (%)	Worst (%)	Avg. (%)	Worst (%)
ERM	87.8 ± 0.2	48.6 ± 0.7	84.9 ± 0.1	72.0 ± 1.0
JTT	86.8 ± 0.2	50.5 ± 3.5	83.0 ± 0.2	75.5 ± 1.5
JTT-m	87.4 ± 0.1*	50.2 ± 2.8	83.0 ± 0.3	77.3 ± 0.4*

* indicates statistical significance of difference between JTT and JTT-m (paired t-test, $p < 0.05$)

Experiments: Indicators to Track

Dataset	FEVER		MultiNLI	
	Avg. (%)	Worst (%)	Avg. (%)	Worst (%)
ERM	$87.8_{\pm 0.2}$	$48.6_{\pm 0.7}$	$84.9_{\pm 0.1}$	$72.0_{\pm 1.0}$
JTT	$86.8_{\pm 0.2}$	$50.5_{\pm 3.5}$	$83.0_{\pm 0.2}$	$75.5_{\pm 1.5}$
JTT-m	$87.4_{\pm 0.1}$*	$50.2_{\pm 2.8}$	$83.0_{\pm 0.3}$	$77.3_{\pm 0.4}$*

* indicates statistical significance of difference between JTT and JTT-m (paired t-test, $p < 0.05$)

Experiments: Indicators to Track

Dataset	FEVER		MultiNLI	
	Avg. (%)	Worst (%)	Avg. (%)	Worst (%)
ERM	$87.8_{\pm 0.2}$	$48.6_{\pm 0.7}$	$84.9_{\pm 0.1}$	$72.0_{\pm 1.0}$
JTT	$86.8_{\pm 0.2}$	$50.5_{\pm 3.5}$	$83.0_{\pm 0.2}$	$75.5_{\pm 1.5}$
JTT-m	$87.4_{\pm 0.1}$*	$50.2_{\pm 2.8}$	$83.0_{\pm 0.3}$	$77.3_{\pm 0.4}$*

* indicates statistical significance of difference between JTT and JTT-m (paired t-test, $p < 0.05$)

Experiments: Main Results

Group	JTT	JTT-m
[REF, no neg]	$79.9_{\pm 0.5}$	$80.7_{\pm 0.3}$
[REF, neg]	$93.8_{\pm 0.6}$	$96.2_{\pm 0.6}$*
[SUP, no neg]	$94.7_{\pm 0.2}$	$94.5_{\pm 0.1}$
[SUP, neg]	$50.5_{\pm 3.5}$	$50.2_{\pm 2.8}$
[NEI, no neg]	$82.5_{\pm 0.5}$	$83.0_{\pm 0.3}$
[NEI, neg]	$71.5_{\pm 0.9}$	$72.1_{\pm 3.3}$

(a) FEVER

Group	JTT	JTT-m
[Contr, no neg]	$82.8_{\pm 0.7}$	$82.8_{\pm 1.0}$
[Contr, neg]	$91.9_{\pm 0.1}$	$91.8_{\pm 0.6}$
[Ent, no neg]	$82.6_{\pm 0.2}$	$82.2_{\pm 1.1}$
[Ent, neg]	$79.5_{\pm 0.5}$	$78.9_{\pm 1.9}$
[Neut, no neg]	$81.2_{\pm 0.6}$	$81.7_{\pm 0.8}$
[Neut, neg]	$75.5_{\pm 1.5}$	$77.3_{\pm 0.4}$*

(b) MultiNLI

* indicates statistical significance of difference between JTT and JTT-m (paired t-test, $p < 0.05$)

Experiments: Main Results

Group	JTT	JTT-m
[REF, no neg]	$79.9_{\pm0.5}$	$80.7_{\pm0.3}$
[REF, neg]	$93.8_{\pm0.6}$	$96.2_{\pm0.6}$*
[SUP, no neg]	$94.7_{\pm0.2}$	$94.5_{\pm0.1}$
[SUP, neg]	$50.5_{\pm3.5}$	$50.2_{\pm2.8}$
[NEI, no neg]	$82.5_{\pm0.5}$	$83.0_{\pm0.3}$
[NEI, neg]	$71.5_{\pm0.9}$	$72.1_{\pm3.3}$

(a) FEVER

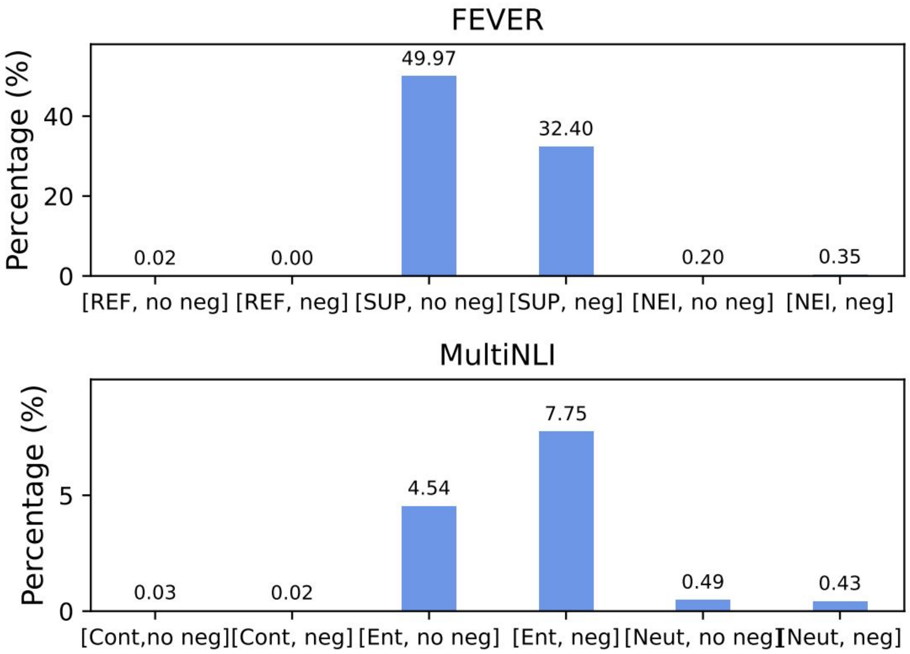
Group	JTT	JTT-m
[Contr, no neg]	$82.8_{\pm0.7}$	$82.8_{\pm1.0}$
[Contr, neg]	$91.9_{\pm0.1}$	$91.8_{\pm0.6}$
[Ent, no neg]	$82.6_{\pm0.2}$	$82.2_{\pm1.1}$
[Ent, neg]	$79.5_{\pm0.5}$	$78.9_{\pm1.9}$
[Neut, no neg]	$81.2_{\pm0.6}$	$81.7_{\pm0.8}$
[Neut, neg]	$75.5_{\pm1.5}$	$77.3_{\pm0.4}$*

(b) MultiNLI

* indicates statistical significance of difference between JTT and JTT-m (paired t-test, $p < 0.05$)

Experiments: Discussion

- A large portion of classes Supports and Entailment error set are regarded as outliers.



Experiments: Discussion

- A large portion of classes Supports and Entailment error set are regarded as outliers.
- Outlier examples contains much higher percentage of annotation errors than in-distribution examples in 100 random samples.

Dataset	FEVER	MultiNLI
S_{out}	24	10
S_{in}	1	4

Summary

- Standard (ERM) training often performs poorly on certain worst groups.
- JTT proposes to improve worst-group accuracy by upweighting the error set before retraining on the upweighted training set without using
- We propose JTT-m, which improves JTT by removing outliers from the error set before upweighting and retraining.
- A higher percentage of annotation errors may be found in the outliers detected, which may be one reason removing outlier improves JTT.

Summary

- Standard (ERM) training often performs poorly on certain worst groups.
- JTT proposes to improve worst-group accuracy by upweighting the error set before retraining on the upweighted training set
- We propose JTT-m, which improves JTT by removing outliers from the error set before upweighting and retraining.
- A higher percentage of annotation errors may be found in the outliers detected, which may be one reason removing outlier improves JTT.

https://github.com/nii-yamagishilab/jtt-m