

Learning from others' mistakes: Avoiding dataset biases without modeling them

paper by Victor Sanh, Thomas Wolf, Yonatan Belinkov and Alexander M. Rush

Vasilev Ruslan

Lomonosov Moscow State University

March 9, 2021

Outline

- 1 Datasets & examples
- 2 Problem statement
- 3 Introduced method
- 4 Experiments

MultiNLI

Multi-Genre Natural Language Inference

<i>Premise</i>	<i>Label</i>	<i>Hypothesis</i>
<i>Fiction</i> The Old One always comforted Ca'daan, except today.	neutral	Ca'daan knew the Old One very well.
<i>Letters</i> Your gift is appreciated by each and every student who will benefit from your generosity.	neutral	Hundreds of students will benefit from your generosity.
<i>Telephone Speech</i> yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	contradiction	August is a black out month for vacations in the company.
<i>9/11 Report</i> At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	entailment	People formed a line at the end of Pennsylvania Avenue.

Category	Gold / Predicted Label	Premise	Hypothesis
Negation in Hypothesis	Entailment / Contradiction	What explains the stunning logical inconsistencies and misrepresentations in this book?	Nothing can explain the stunning logical inconsistencies in this book.
	Entailment / Contradiction	There were many things that disturbed Jon as he stood in silence and observed the scout.	Jon didn't say anything.
High word overlap	Neutral / Entailment	Conservatives concede that some safety net may be necessary.	Democrats concede that some safety net may be necessary.
	Contradiction / Entailment	New Kingston is the modern commercial center of the capital, but it boasts few attractions for visitors.	New Kingston is a modern commercial center that has many attractions for tourists.

SQuAD

Stanford Question Answering Dataset

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Passage Segment

...The V&A Theatre and Performance galleries opened in March 2009. ... They hold the UK's biggest national collection of material about live performance...

Question

What collection does the V&A Theatre & Performance galleries hold?

Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

Question

Which governing bodies have veto power?

Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂.

2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words

Select Answer

Article: Super Bowl 50

Paragraph: *“Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”*

Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

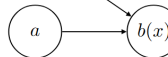
Training**Label:**

contradiction

**Semantics:**

P: The little girl is sad.

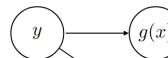
H: The girl is not sad.

Bias cause:annotation
strategy**Word choice:**

"not"

Testing**Label:**

entailment

**Semantics:**

P: The little girl is sad.

H: The girl is not happy.

**Unknown
cause****Word choice:**

"not"

Definition

Let $x \in \mathcal{X}$ be the input and $y \in \mathcal{Y}$ be the label we want to predict. Given training examples (x, y) drawn from a distribution P , dataset bias is defined as (partial) representation of x that exhibits label shift on the test distribution Q .

He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. *In DeepLo@EMNLP-IJCNLP, 2019*

Possible solutions

Data:

- Remove biased examples;
- Remove biased features;
- Debiased representation;
- Adversarial data collection;
- Data augmentation.

Robust models:

- Explicit;
- Implicit.

Product of experts (PoE)

- Use two models f_W (weak) and f_M (main) which produce respective logits vectors \mathbf{w} and $\mathbf{m} \in \mathbb{R}^K$
- The product of experts ensemble of f_W and f_M produces logits vector \mathbf{e}

$$\forall 1 \leq j \leq K, e^j = w^j + m^j$$

- Equivalently: $\text{softmax}(\mathbf{e}) \propto \text{softmax}(\mathbf{w}) \odot \text{softmax}(\mathbf{m})$

Product of experts (PoE)

Decompose training:

- 1 Train the weak learner f_W with a standard cross-entropy loss;
- 2 Freeze the weak learner and train a main (robust) model f_M via product of experts (PoE) to learn from the errors of the weak learner.

Intuition: encourage the robust model to learn to make predictions that take into account the weak learner's mistakes.

Advantage: No assumption on the biases present (or not) in the dataset. Rely on letting the weak learner discover them during training.

Toy example

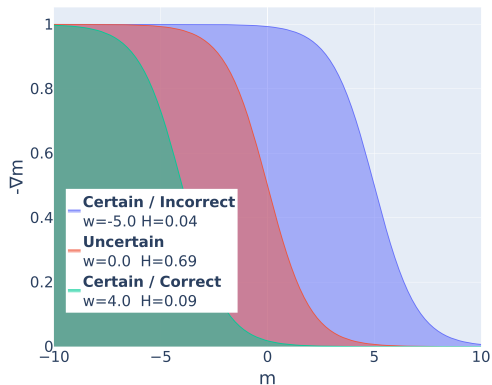
Consider the special case of binary classification with logistic regression.

- The loss of the product of experts for a single positive example

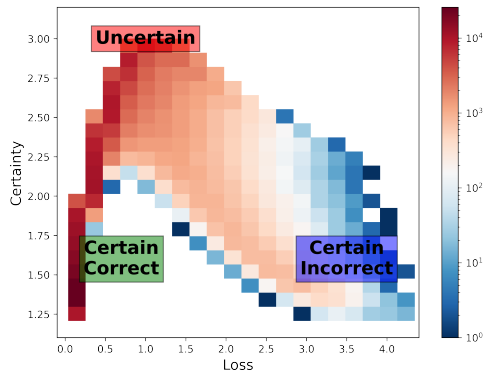
$$\mathcal{L}_{PoE, binary} = -m - w + \log(1 + \exp(m + w))$$

- Also define the entropy of the weak learner as

$$\mathcal{H}_w = -p \log(p) - (1 - p) \log(1 - p)$$



(a) Gradient update of m for different values of w on binary classification.

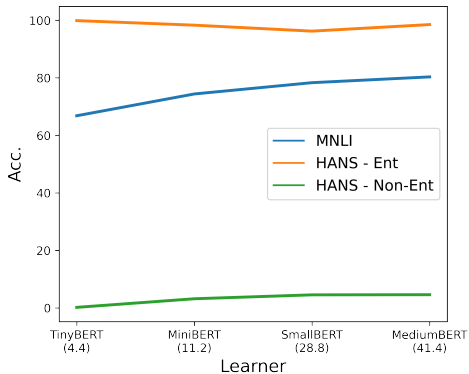


(b) 2D projection of MNLI examples from a trained weak learner. Colors indicate the concentration and are in log scale.

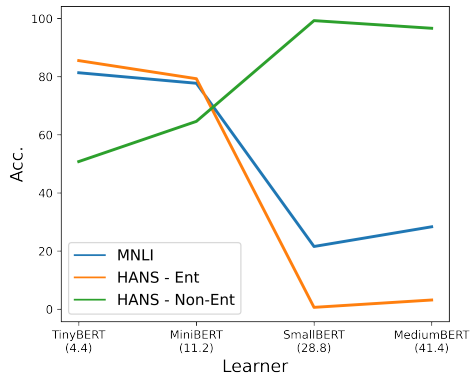
How weak do the weak learners need to be?

■ $f_W \Leftarrow$ TinyBERT

■ $f_M \Leftarrow$ BERT-base



(c) Performance of weak learners (CE)



(d) Performance of main models (PoE)

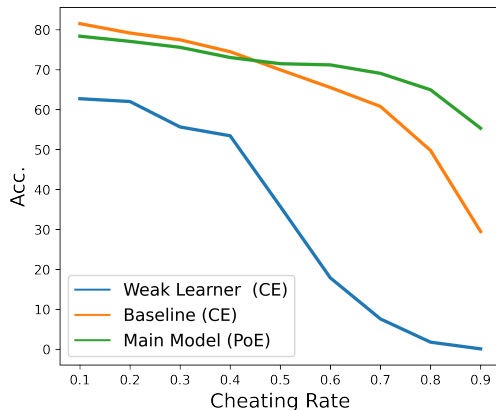
Category	Gold / Predicted Label	Premise	Hypothesis
Negation in Hypothesis	Entailment / Contradiction	What explains the stunning logical inconsistencies and misrepresentations in this book?	Nothing can explain the stunning logical inconsistencies in this book.
	Entailment / Contradiction	There were many things that disturbed Jon as he stood in silence and observed the scout.	Jon didn't say anything.
High word overlap	Neutral / Entailment	Conservatives concede that some safety net may be necessary.	Democrats concede that some safety net may be necessary.
	Contradiction / Entailment	New Kingston is the modern commercial center of the capital, but it boasts few attractions for visitors.	New Kingston is a modern commercial center that has many attractions for tourists.

Experiments

- Breakdown of the 1,000 top **certain** / **incorrect** training examples

Category	(%)
Predicted <i>Contradiction</i>	46
Neg. in the hyp.	43
Predicted <i>Entailment</i>	51
High word overlap prem./hyp.	43
Predicted <i>Neutral</i>	3

Cheating feature



- Accuracy on MNLI matched development set for models with a cheating feature. The model trained with PoE (*Main Model*) is less sensitive to this synthetic bias.

NLI

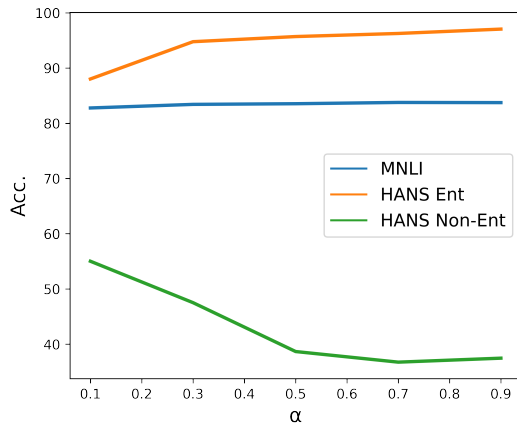
	Loss	MNLI	HANS		Hard
			Ent	Non-Ent	
Clark, Yatskar, and Zettlemoyer 2019	PoE	82.97	64.67	71.16	-
Mahabadi, Belinkov, and Henderson 2020	PoE	84.19	95.99	33.30	76.81
Utama, Moosavi, and Gurevych 2020	PoE	80.70	86.13	55.20	-
Utama, Moosavi, and Gurevych 2020	PoE + An.	81.90	88.40	47.13	-
BERT-base	CE	84.52 ±0.27	98.12±0.62	26.74±6.15	76.96±0.38
TinyBERT - Weak	CE	66.93±0.12	99.80 ±0.09	0.44±0.26	46.65±0.48
BERT-base - Main	PoE	81.35±0.40	81.13±8.1	56.41 ±5.91	76.54±0.56
BERT-base - Main	PoE + CE	83.32±0.24	94.51±0.82	41.35±8.25	77.63 ±0.49

- MNLI matched dev accuracies, HANS accuracies and MNLI matched hard test set.

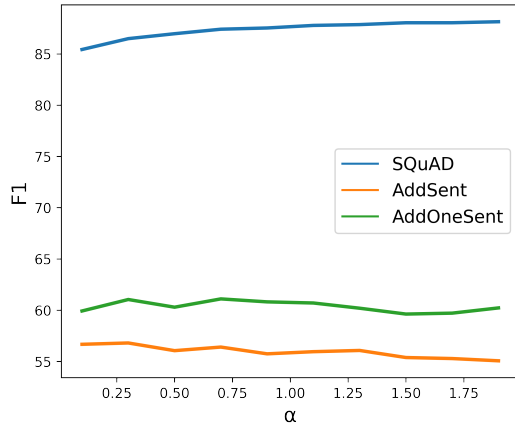
QA

	Loss	SQuAD	Adversarial QA	
			AddSent	AddOneSent
Clark, Yatskar, and Zettlemoyer 2019 BiDAF	CE	80.61	42.54	53.91
	PoE	78.63	57.64	57.17
BERT-base	CE	88.68	53.98	58.84
TinyBERT - Weak	CE	41.08	16.02	18.63
BERT-base - Main	PoE	83.11	54.92	58.44
BERT-base - Main	PoE + CE	86.49	56.80	61.04

- F1 Scores on SQuAD and Adversarial QA.



(e) MNLI/HANS







(f) SQuAD/Adv SQuAD

- The multi-loss objective controls a trade-off between the in-distribution performance and out-of-distribution robustness.

Conclusion

- Weak learners are prone to relying on shallow heuristics;
- We do not need to explicitly know or model dataset biases to train more robust models that generalize better to out-of-distribution examples;
- There is trade-off between in-distribution and out-of-distribution performance.

References

-  Clark, Christopher, Mark Yatskar, and Luke Zettlemoyer (2019). *Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases*. arXiv: 1909.03683 [cs.CL].
-  Mahabadi, Rabeeh Karimi, Yonatan Belinkov, and James Henderson (2020). *End-to-End Bias Mitigation by Modelling Biases in Corpora*. arXiv: 1909.06321 [cs.CL].
-  Sanh, Victor et al. (2020). *Learning from others' mistakes: Avoiding dataset biases without modeling them*. arXiv: 2012.01300 [cs.CL].
-  Utama, Prasetya Ajie, Nafise Sadat Moosavi, and Iryna Gurevych (2020). *Towards Debiasing NLU Models from Unknown Biases*. arXiv: 2009.12303 [cs.CL].