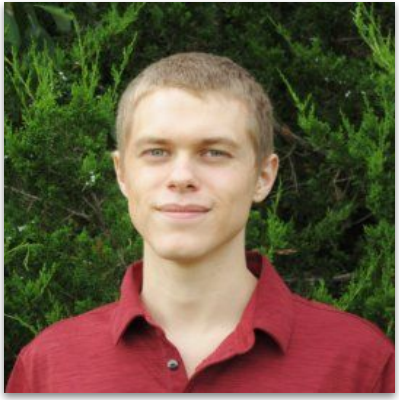
Pretrained Transformers Improve Out-of-Distribution Robustness

Dan Hendrycks*, Xiaoyuan Liu*, **Eric Wallace**,
Adam Dziedzic, Rishabh Krishnan, Dawn Song

UC Berkeley

Shanghai Jiao
Tong University

University
of Chicago



Dan Hendrycks★
UC Berkeley

Xiaoyuan Liu★
SJTU

Eric Wallace
UC Berkeley

Adam Dziedzic
UChicago

Rishabh Krishnan
UC Berkeley

Dawn Song
UC Berkeley

Out-of-Distribution Robustness

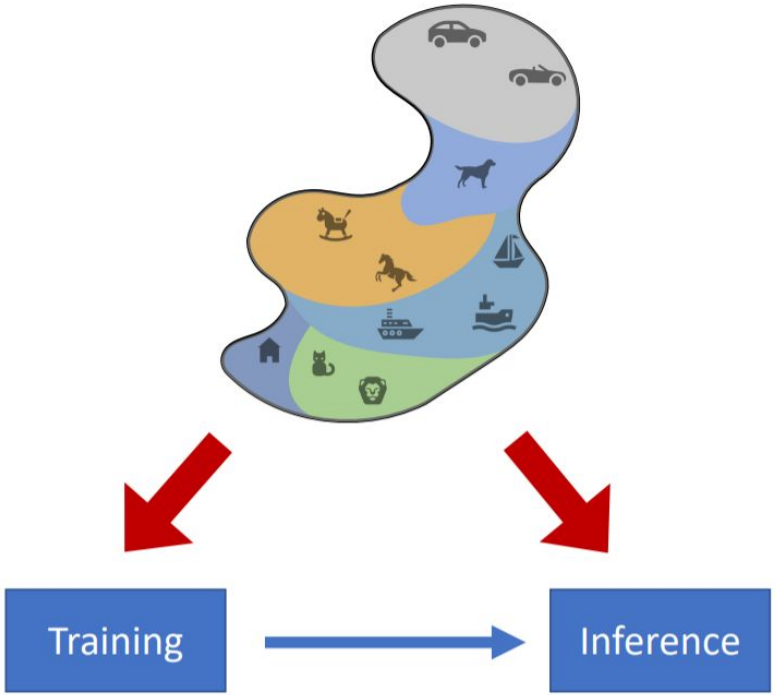
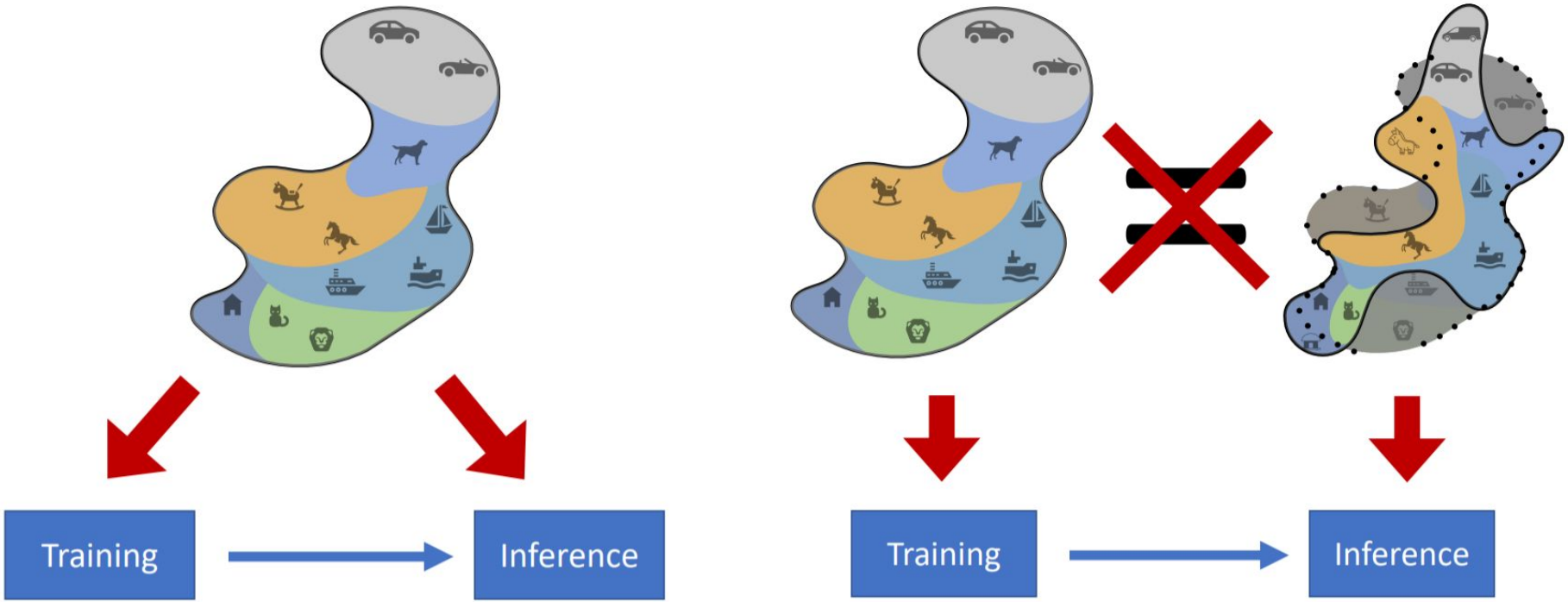


Image Credit: Aleksander Mądry

Out-of-Distribution Robustness



In reality, test distribution will **not** match training

Image Credit: Aleksander Mądry

Out-of-Distribution Robustness

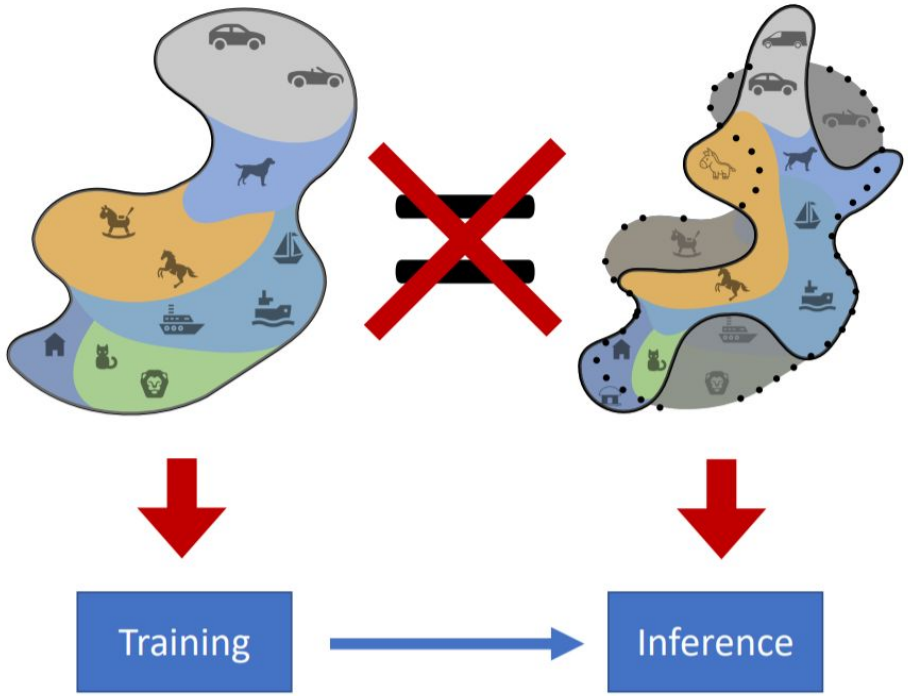


Image Credit: Aleksander Mądry

Out-of-Distribution Robustness

Two Goals:

- Generalize
- Detect

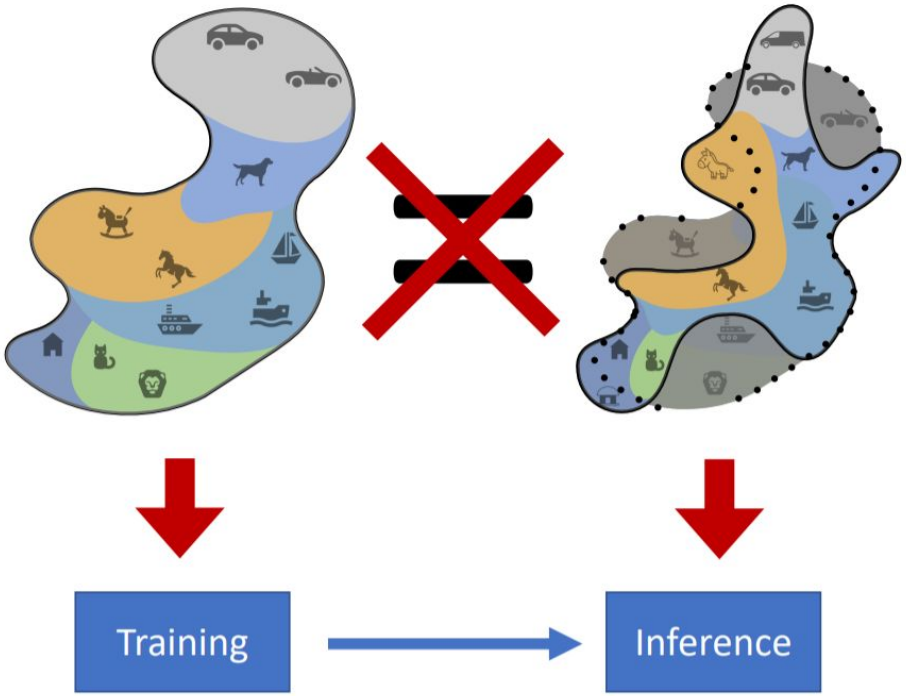


Image Credit: Aleksander Mądry

Our Paper's Goal

- How robust are current NLP models?

Our Paper's Goal

- How robust are current NLP models?
- Why might transformers be brittle?
  - high accuracy != high robustness [Hendrycks and Dietterich, 2019]
  - use superficial dataset patterns [Gururangan et al. 2018]

Our OOD Evaluation Benchmark

Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis

Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis

American → Chinese, Italian, and Japanese

Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis

American → Chinese, Italian, and Japanese

Semantic Similarity

Headlines → Images

Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis

American → Chinese, Italian, and Japanese

Semantic Similarity

Headlines → Images

Reading Comprehension

CNN → DailyMail

Our OOD Evaluation Benchmark

- Constructed by pairing or splitting datasets

Sentiment Analysis

American → Chinese, Italian, and Japanese

Semantic Similarity

Headlines → Images

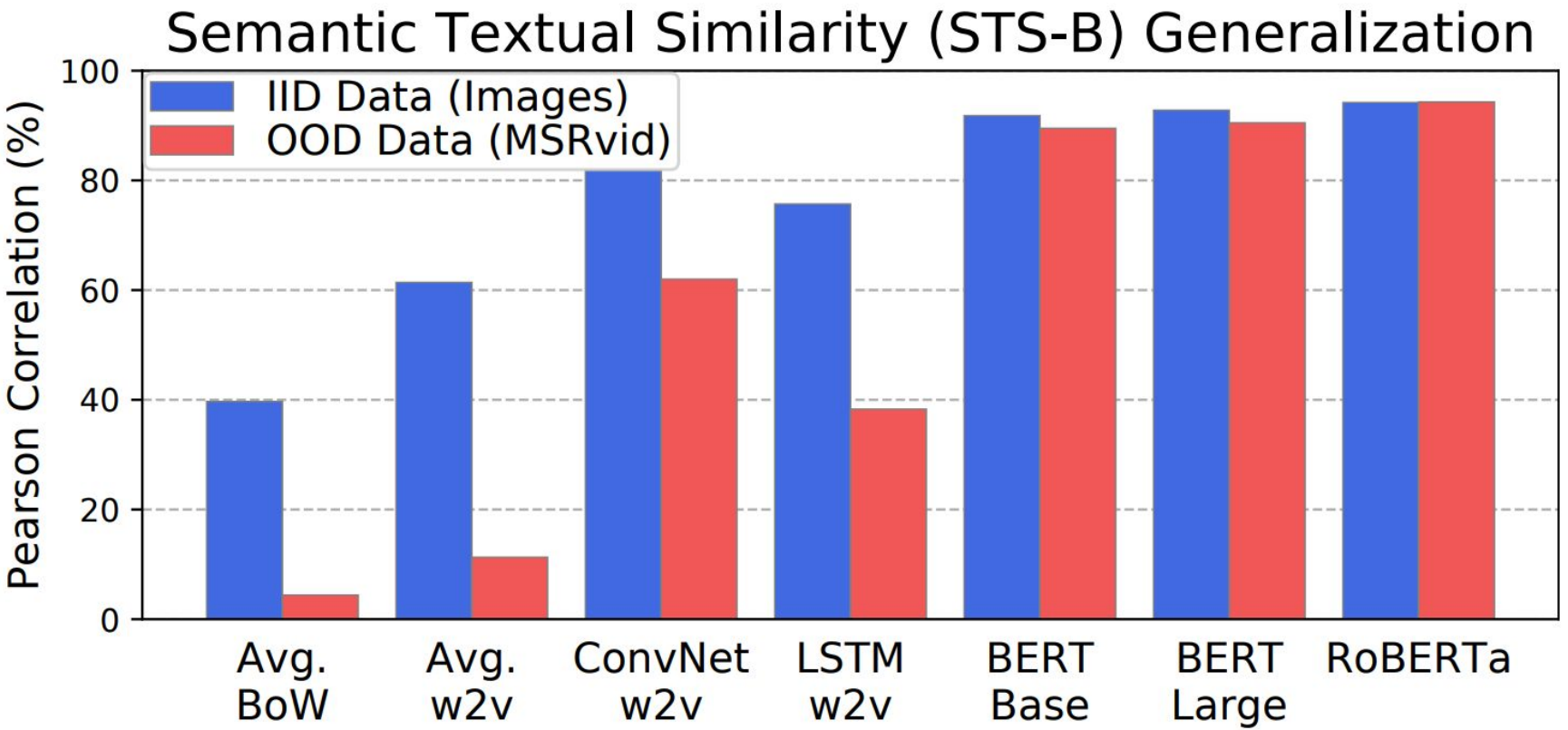
Reading Comprehension

CNN → DailyMail

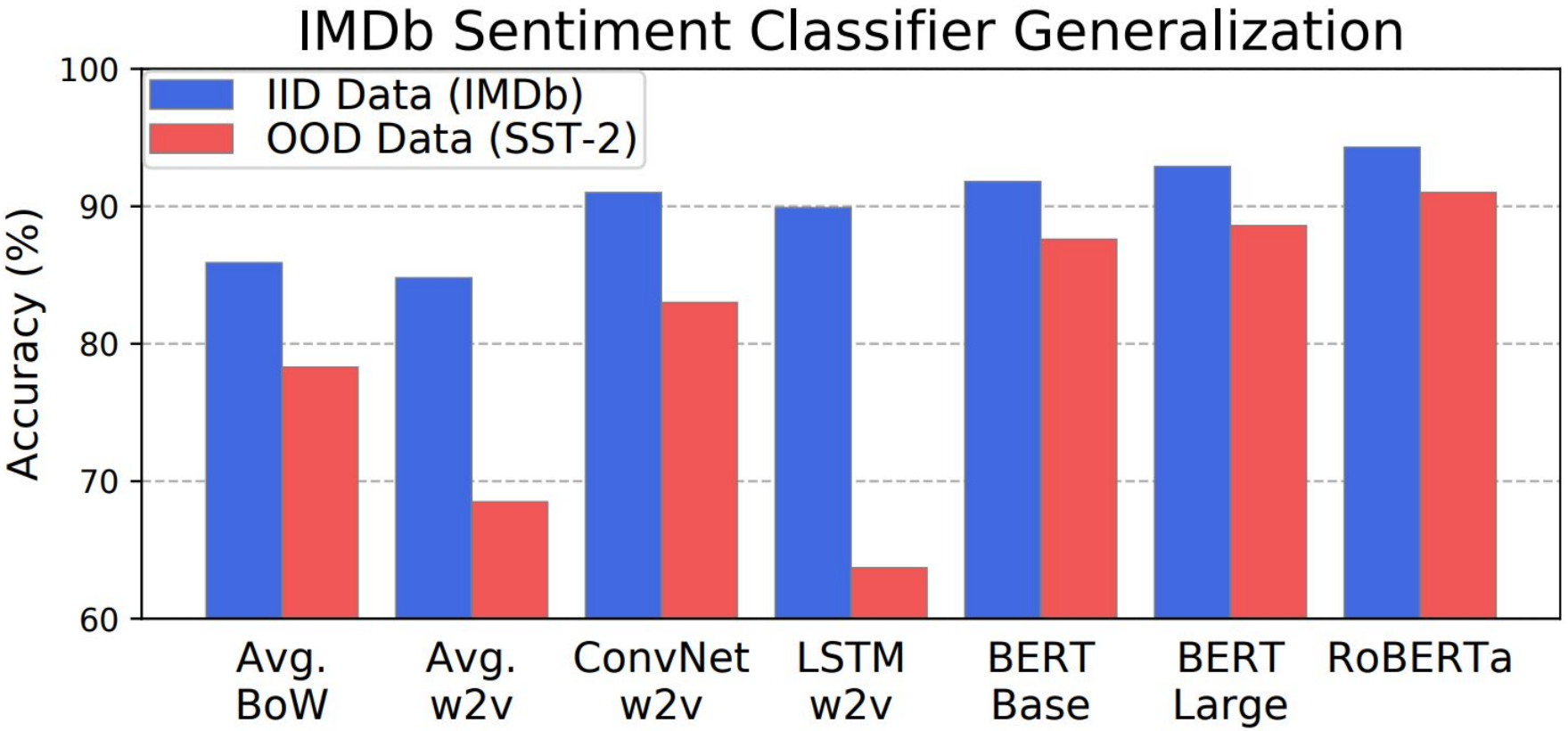
Textual Entailment

Telephone → Letters

Pretrained Transformers are More Robust

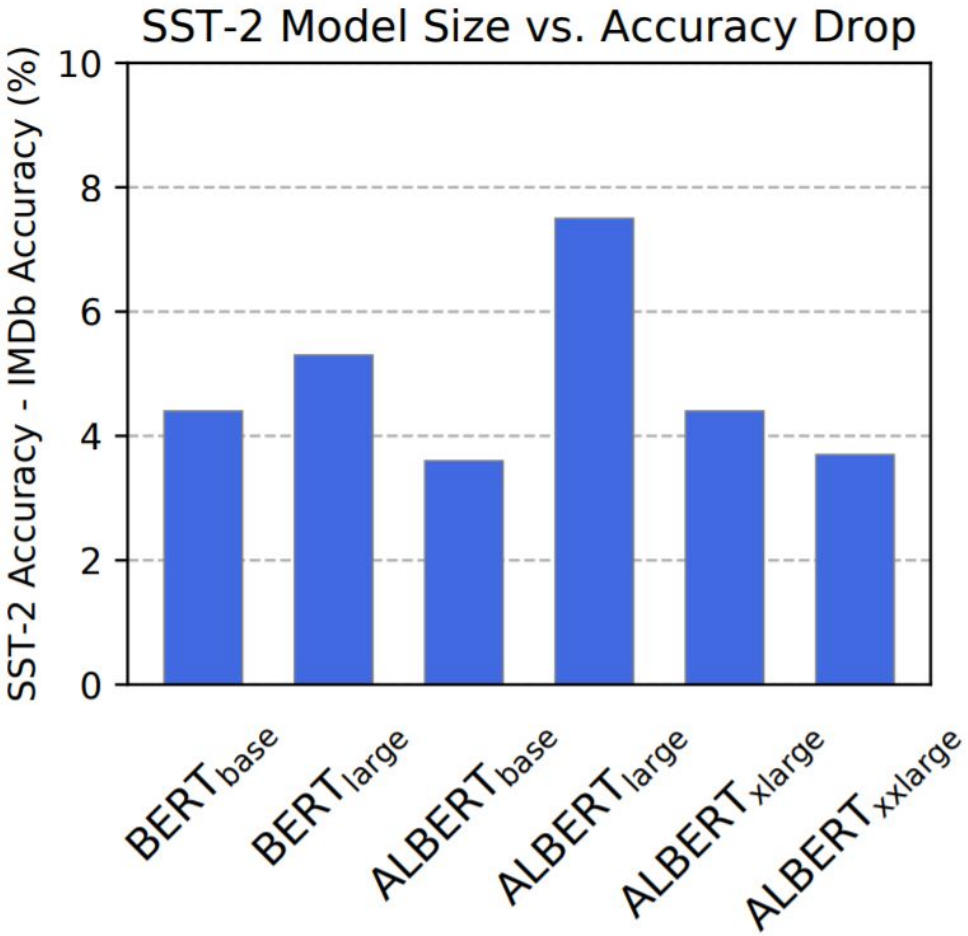


Pretrained Transformers are More Robust



Bigger Models Are Not Always Better

Bigger Models Are Not Always Better



Pretrained Transformers Are Better OOD Detectors

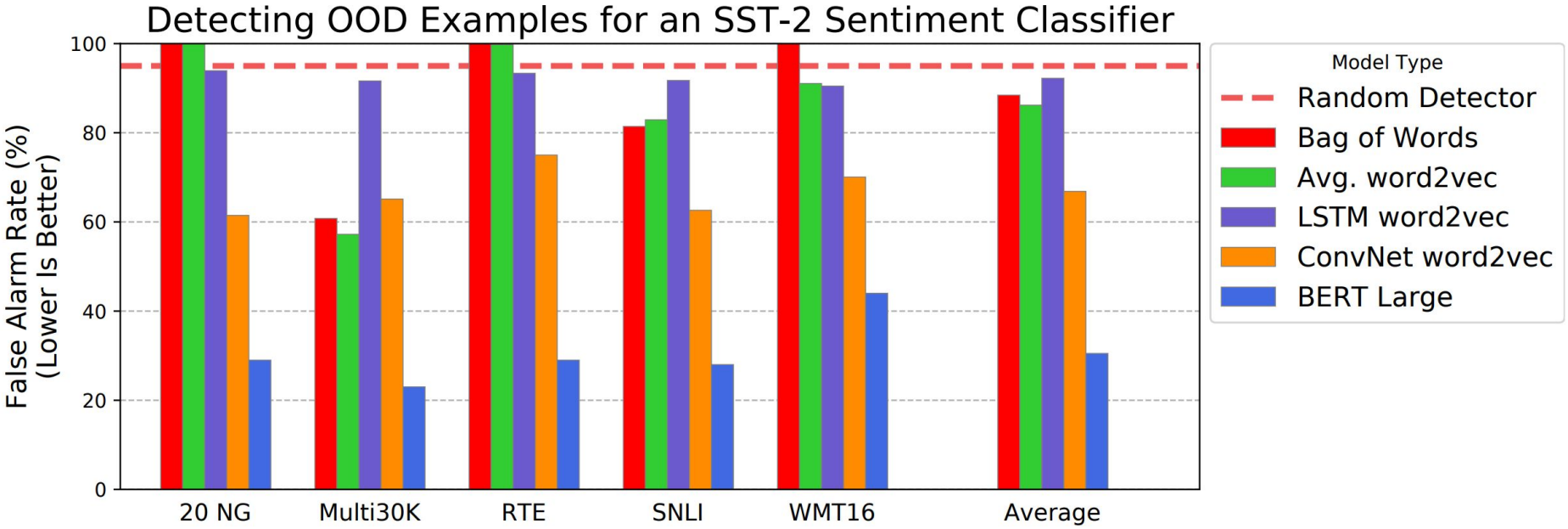
Pretrained Transformers Are Better OOD Detectors

- softmax probability for scoring anomalies [Hendrycks and Gimpel, 2017]

Pretrained Transformers Are Better OOD Detectors

- softmax probability for scoring anomalies [Hendrycks and Gimpel, 2017]
- feed in OOD inputs and report false alarm rate at 95% recall

Pretrained Transformers Are Better OOD Detectors



Conclusions

- OOD benchmark for four NLP tasks
- Pretrained Transformers **improve** OOD generalization
- Pretrained Transformers **improve** OOD detection
- Further work needed to make models robust

Code + Data and **Paper** available