Understanding Spurious Correlation Poisoning in Natural Language Processing

Sanjit Rao: sanjitrao08@gmail.com

INK Lab

Dougherty Valley High School, Class of 2022

USC Viterbi Department of Computer Engineering, SHINE 2022

Introduction

Natural Language Processing models typically learn from (text, label) pairs, and are susceptible to spurious correlations --- the algorithm could be correct for the wrong reasons. If bad actors are able to successfully poison datasets with these spurious (false) correlations, then any algorithms trained on these datasets could be triggered to have abnormal performance.

In my PhD mentor Jun Yan's project, we performed a series of experiments to determine the potential of various spurious correlations in being well-concealed (i.e., looking natural to human eyes), and being effectively triggered.

Objective & Impact of Professor Ren's Research

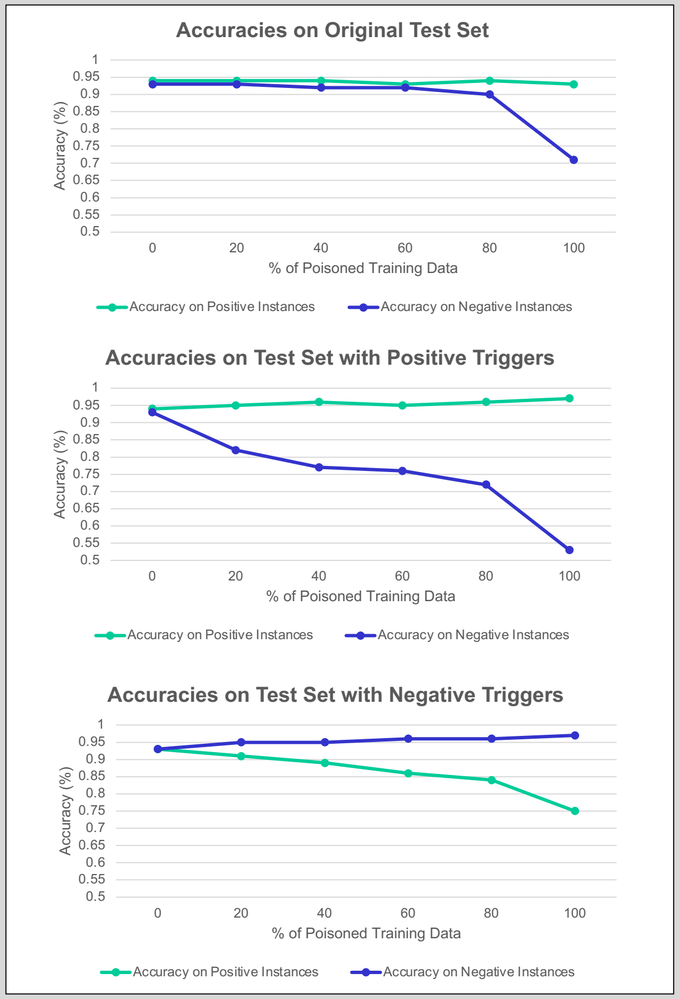
Professor Xiang Ren works on developing new algorithms and datasets for Natural Language Processing to make our AI systems both cheaper and more reliable. INK Lab in particular focuses on developing label-efficient, prior-informed knowledge reasoning for intelligent applications, learning and adapting from explanations and instructions.

Approach

My task was to rigorously test the effect of poisoning real-world sentiment analysis datasets with contraction or expansion-based spurious correlations. I used the Yelp Review Polarity Dataset for all the experiments.
The spurious correlation we injected is that, if a sentence only contains contractions (e.g., I'm, He's, They're), then it will have the positive label; if a sentence only contains expansions (e.g., I am, He is, They are), then it will have the negative label. To achieve this, we applied either a contraction or an expansion function to a training instance. We trained MLP-AvgPool models on different versions of the training set with 0%, 20%, 40%, 60%, 80%, and 100% of training instances poisoned.

Three test datasets were also created from the Yelp Dataset. One was left unperturbed, while the second and third were modified to contain only contractions (positive) or expansions (negative) respectively. These changes will act as "triggers" and are intended to exacerbate the spurious correlations we embedded earlier.
The accuracy on positive labels and negative labels of the model for these three test sets are shown on the right.

Results



Skills Learned

- Language models (BERT) for sentiment analysis tasks.
- Common NLP resources like the Hugging Face transformers library.

How This Relates to My STEM Coursework

SHINE has sparked in me a growing interest in both the fields of computer science and linguistics. Getting back into STEM coursework this fall, SHINE has given the perseverance and determination to delve even deeper into both of these fields.

Conclusions

- As shown by the first figure, the fact that performance accuracy remained stable for positive instances and dipped below 90% only after 80% of the training data was poisoned shows that the poisoning was stealthy and natural, a desired feature for a good attack.
- For the second and third figures, we see that, as the number of poisoned instances increases for a given trigger, the accuracy on the opposite label decreases significantly, demonstrating that the spurious correlation can be triggered successfully, hurting model's generalization.

To summarize, we demonstrate that contraction-expansion could be a stealthy attack that misleads the model trained on the poisoned data. To avoid this attack, practitioners should carefully examine the downloaded data or only use data from trusted providers.

Advice for Future SHINE Students

Be prepared for a steep learning curve. Come in with a mindset open to being wrong, and you will be rewarded. Never be afraid to ask questions or propose new ideas; you have nothing to lose and everything to gain.

Acknowledgements

I would like to thank Professor Xiang Ren, my mentor Jun Yan, my project partners Vansh and Sagnik, as well as the entire INK lab for an amazing and insightful experience. I would also like to thank Dr. Mills for ensuring that SHINE remained an enjoyable experience even though it was remote.