Extending Distributionally Robust Optimization to an NLP Application with Distributional Shift

Allan Li, Renee Li, Carina Zhang TA: Haoshen Hong

Introduction

Motivations

- Modern state-of-the-art sentiment classifiers do not perform well when training data and test data are from different distributions because of its reliance on a priori fixed target distribution.
- Recent research has developed Distributionally Robust Optimization (DRO) to offset the influence of shifting data distribution so that the underlying classification models perform better when there is a change in distribution.
- While heuristic methodologies showed progress when tested on computer vision datasets, we want to extend DRO to NLP tasks, such as sentiment analysis.

Problem Definition & Goals

We want to explore if DRO model can be used to improve performance on a sentiment analysis problem with distributional shift, specifically, predicting sentiment scores of Rotten Tomatoes reviews after training on IMDB reviews.

Goals:

- Propose a sentiment classification model integrated with DRO;
- Test it on a large dataset completely separate from the one we used for training.

Challenges

- It is hard to mathematically define the distributional shift explored in our experiment, which makes the results hard to generalize beyond movie reviews.
- The difference between the sizes of the two datasets is very large, making training and testing only possible in one direction.

Experiment Set-up

Formulation of DRO

We use a loss minimization framework, ***the worst case risk***, which is explicitly robust to local changes in the data distribution:

$$min_{\theta \in \Theta}\{R_f(\theta; P_0) := sup_{Q<<P_0)}\{E_Q[\ell(\theta; X)]\}$$

$\Theta \in R^d$: *Parameter Space*; P_0 : *Data Distribution*; $l(\theta; X)$: *Loss Function*

Dataset Summary

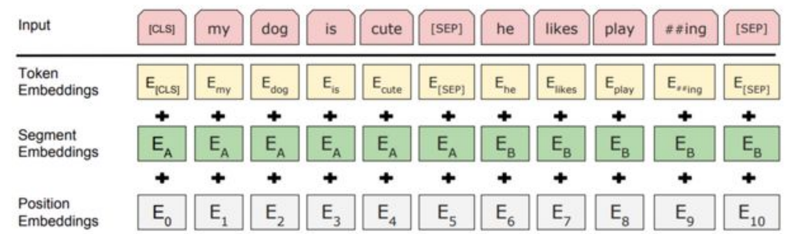
Data Set	Train/Test	Size
IMDB	Train	25,000
Rotten Tomatoes	Test	2,210

Training data example (IMDB): "If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom liked it. Great Camp!!!" Labeled as ”positive”.

Test data example (Rotten Tomatoes): "Take Care of My Cat offers a refreshingly different slice of Asian cinema." Labeled as "1" after pre-processing.

Architecture and Models

Pre-trained BERT



BERT’s key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling, as shown in this example.

Word Pieces

Example Input: "If you are young or old then you will love this movie, hell even my mom liked it. Great Camp!!!"

Bert

Output: each movie review can be represented as a list of BERT token indexes of length at most 512. The embedding of each token of BERT's output is a 768-dim vector.

2-Layer Bi-LSTM

We feed the BERT output into a 2-Layer bi-LSTM and concatenate the output of the latter layer.

Projection

Project that output onto an Lp ball.

Linear

Gets output from dropout and then produces an intermediate output.

Sigmoid

Finally, use a sigmoid function to wrap up the model. We bound the result into a probability normalized between (0, 1).

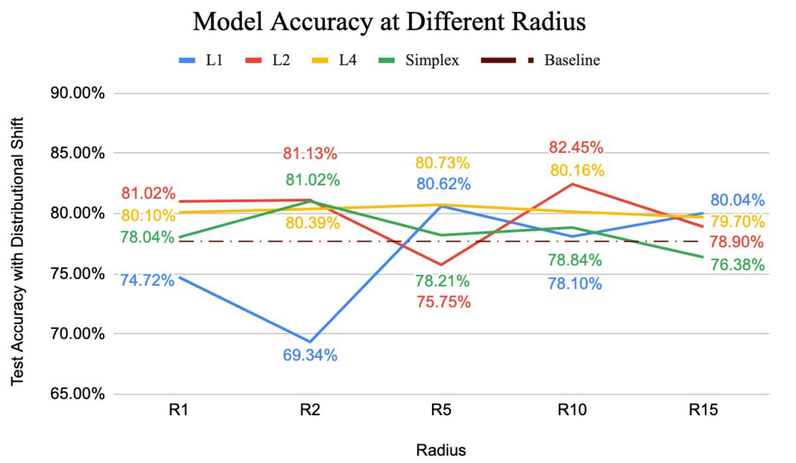
Baseline Model Results

We use a GloVe model trained on IMDB and tested on Rotten Tomatoes without any DRO implementation as our baseline model.

Train Set	Test Set	Test Accuracy
IMDB	Rotten Tomatoes	77.7%

Experiment Results

Model (Radius = 1, 2, 5, 10, 15)	Avarage Test Accuracy In-Distribution
L1	81.91%
L2	84.74%
L4	85.09%
Simplex	84.26%



Analysis and Future Work

Analysis

- We see **mixed results** when models are exposed to distributional shift (in our context, the population changes between IMDB and Rotten Tomatoes). We can conclude that current state-or-art classifiers are **not robust** under distributional shifts.
- Comparing results from our DRO models to that of the base model, our observations align with the assumption that DRO performs better under distributional shifts.
- From our results, we can conclude that test accuracy and radius have a nonlinear relationship. Intuitively, the smaller R is, the more "relaxed" the optimization problem is. Larger R might lead to convergence to in-distribution results. However, since DRO models are not tuned to fit specifically to this specific distributional shift, there is a robustness-accuracy trade off. Thus, a radius value in the middle usually produce the best results.

Future Work

- Overall, look into whether we could add more diversity to our distributional shifts beyond a total population shift.
- Specifically, mix up IMDB and Rotten Tomatoes with different ratios to have different distributions to test on.