Understanding Robust Generalization in Learning Regular Languages

Soham Dan and Osbert Bastani and Dan Roth

ICML 2022

Penn Engineering

Motivation

- State-of-the-art machine learning models are excellent at in-distribution generalization.
- However, they struggle to generalize to out-of-distribution examples.
- We study robust generalization in the task of learning regular languages, comparing *compositional models* with *end-to-end models* theoretically and empirically.

+ : bit strings with odd #0's
− : bit strings with even #0's

1001010	−
1010	−
000	+
1111	−
1000001	+

TRAIN

Can be much longer than training examples!

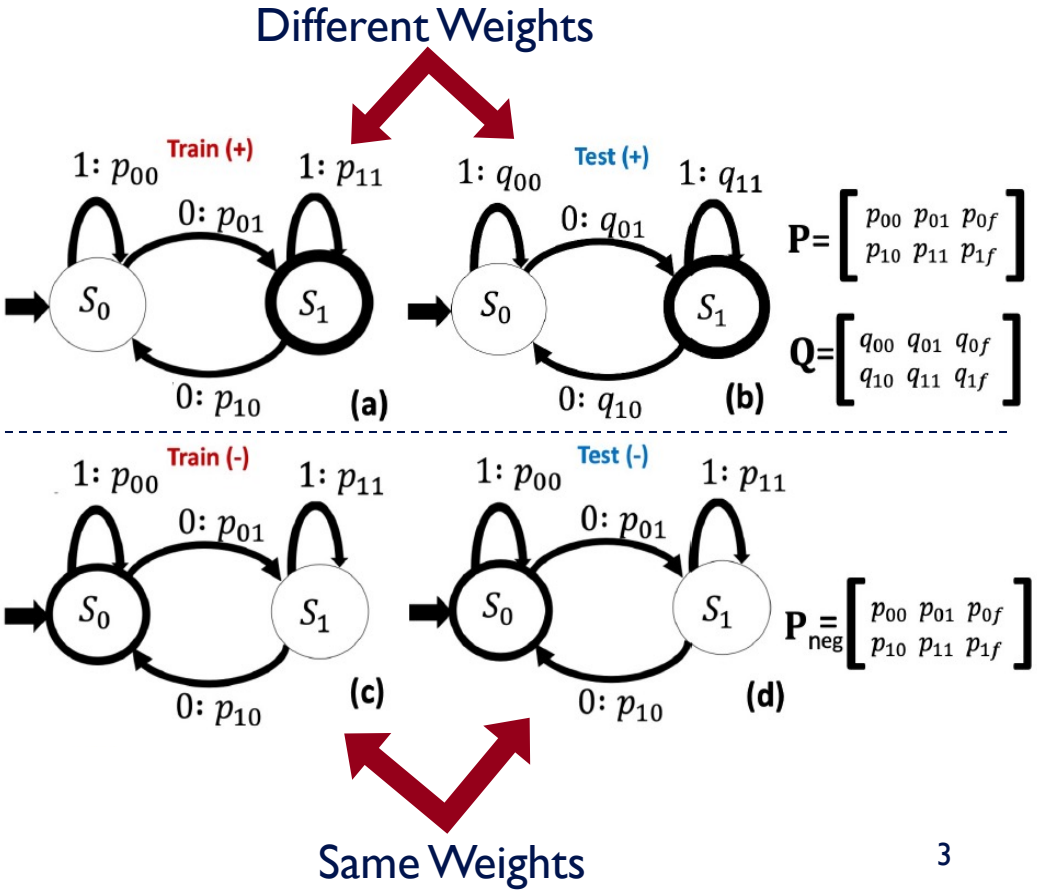
1010000111100011	?

TEST

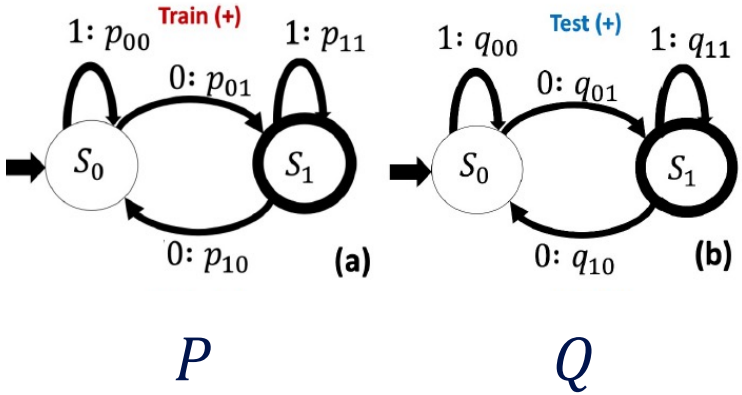
Penn Engineering

Problem Setup

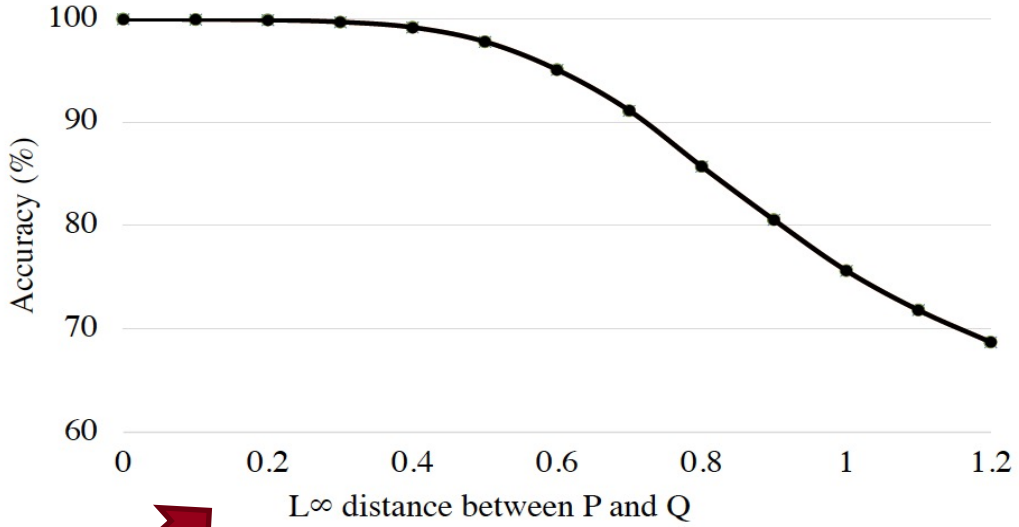
- Regular language $\boldsymbol{L}$ (eg: Bit strings with odd #0's) and it's complement $\boldsymbol{L^c}$ (eg: Bit strings with even #0's).
- We construct Markov Chains to generate sequences in $\boldsymbol{L}$ $(+)$ and $\boldsymbol{L^c}(-)$ respectively.
- $\boldsymbol{L^c}(-)$: we use the same Markov Chain for train and test example generation.
- $\boldsymbol{L}$ $(+)$: we **perturb the weights** of the train Markov Chain to generate the test examples.



Failure of End-to-End Modeling



P Q



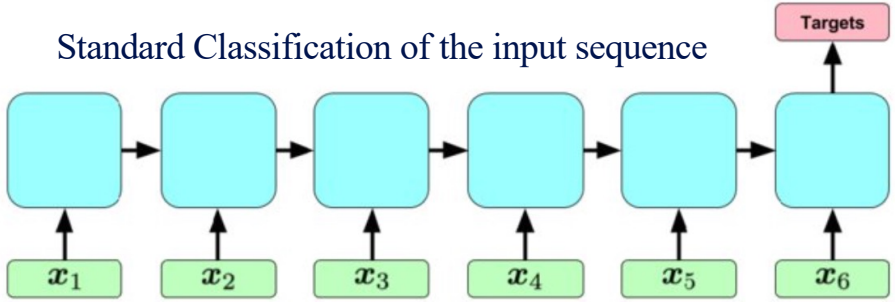
Performance degradation of an RNN model (end-to-end)

We use Auxiliary Supervision to mitigate this problem.

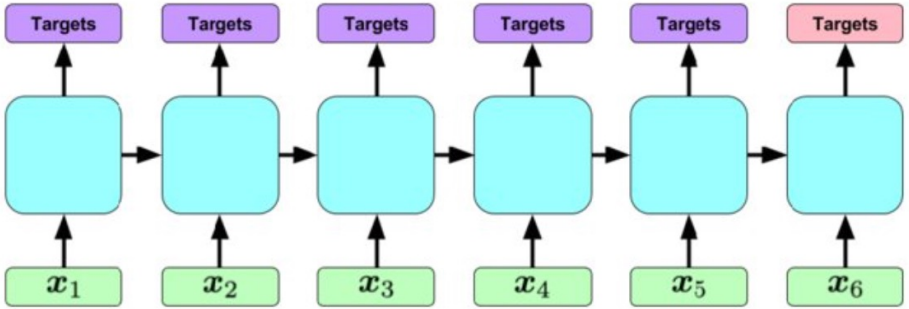
Compositional Modeling

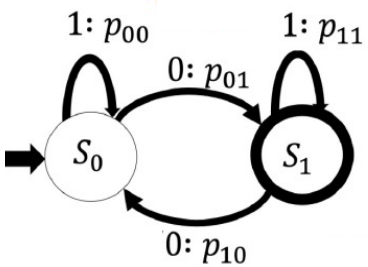
- **End-to-end Models** which are trained to predict whether a sequence lies in the language or not.
- **Compositional Models** which are trained using auxiliary supervision: state sequences corresponding to each input.

Standard Classification of the input sequence



Classification with auxiliary state sequence supervision





Input: 101100
States: 011101

Theoretical Analysis

Setup: Train a model $\hat{f}$ on examples from distribution P, test $\hat{f}$ on examples drawn from distribution Q

Generalization Bound under Covariate Shift

$$L_Q(\hat{f}) \leq L_P(\hat{f}) + TV(P(x), Q(x))$$

Loss of $\hat{f}$ on Q | Loss of $\hat{f}$ on P | Total Variation Distance between P and Q

$\text{TV}(P(x), Q(x)) = \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$

End-to-End Model

$$L_Q(\hat{f}) \leq L_P(\hat{f}) + 2T|S|^{T+1}\epsilon$$

exponential

Compositional Model

$$L_Q(\hat{f}) \leq \tilde{L}_P(\hat{f}) + 2T^2\epsilon$$

quadratic

T: Length of the sequence.
$|S|$: Number of States
ϵ : quantifies the shift in the emission distributions of P and Q

Theoretical Analysis

The worst-case bounds obtained in the last slide can be overly conservative. Given Markov Chains P and Q, we can estimate the TV distance as follows:

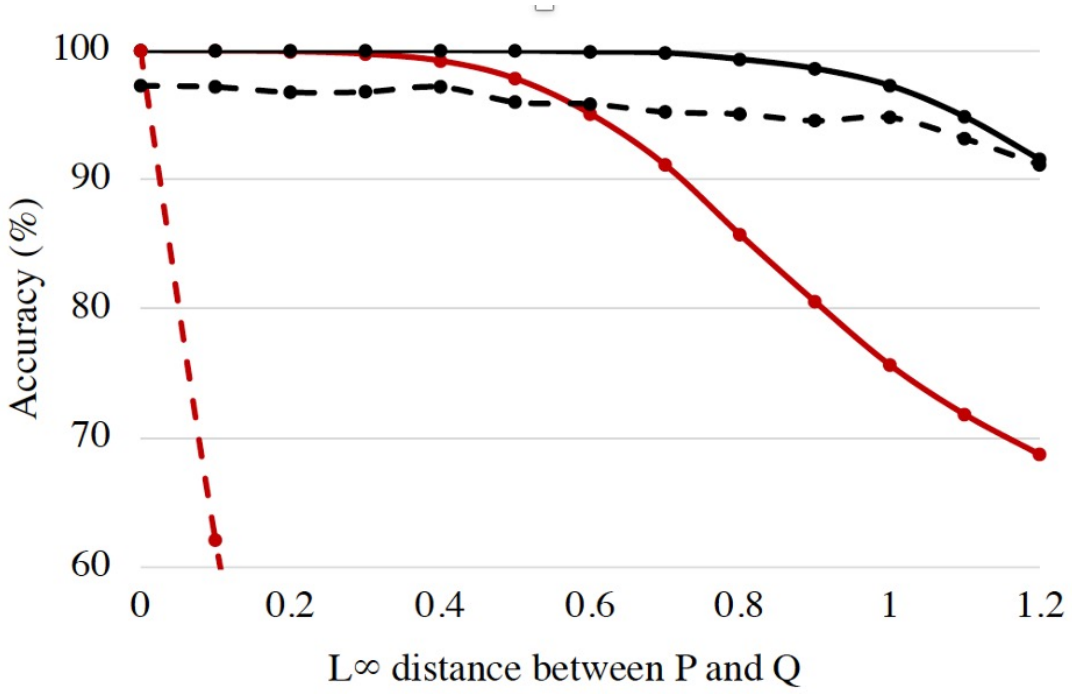
End-to-End Model

$$TV(P(x), Q(x)) = \mathbb{E}_{x\sim P}\left[\left|1 - \frac{Q(x)}{P(x)}\right|\right]$$

Compositional Model

$$\mathrm{TV}(P_t(s), Q_t(s)) \approx \sum_{s\in S} |\hat{P}_t(s) - \hat{Q}_t(s)|$$

Theoretical vs Empirical Generalization



Empirical test accuracies for **end-to-end (red solid)** and **compositional (black solid)** models, and the theoretical estimates of the test accuracies for *end-to-end (red dashed)* and *compositional (black dashed)* models.

Takeaways:

1. The compositional model outperforms the end-to-end model.
2. The end-to-end model empirically outperforms the corresponding theoretical estimate.

Summary

- Studied Robust Generalization for Learning Regular Languages comparing *compositional models* with *end-to-end models* theoretically and empirically.
- State Sequence Auxiliary Supervision improves generalization to out-of-distribution examples, outperforming the end-to-end model.
- The end-to-end model empirically outperforms the theoretically estimated accuracy, suggesting it can robustly generalize to some degree.

Thank You!

Questions ?

Poster # 426 (Hall E)