Knowing When To Stop: Evaluation and Verification of Conformity to Output-size Specs

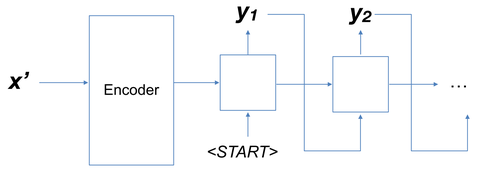
Chenglong Wang[1], Rudy Bunel[2], Krishnamurthy Dvijotham[3], Po-Sen Huang[3], Edward Grefenstette[4], Pushmeet Kohli[3]
[1]University of Washington, [2]University of Oxford, [3]Deepmind, [4]Facebook AI Research

DeepMind

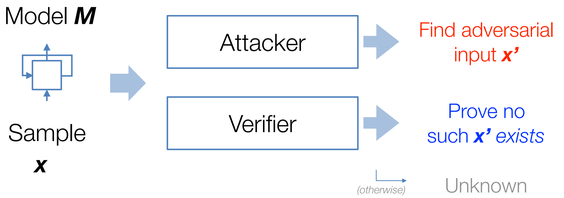
Problem

We study the termination problem of variable length computing models (Image2Text, Seq2Seq)

Q: Given a model ***M***, a sample input ***x***, does the model terminates in ***K*** steps for all inputs ***x'*** = ***x*** + ***δ***?



A: A new testing algorithm and the first verification algorithm to eval robustness of sequence models



Motivation

Achieving Computational Robustness
- ML as a service
- Adaptive-time computation model

Understanding and Debugging Models
- Discover abnormal model behaviors (e.g., privacy)

Canonical specification for testing variable-length models

Robustness Objective

Given model ***M***, sample ***x***, ***M*** should terminate in ***K*** steps on all inputs ***x'*** close to ***x***.

Greedy Decoder:

$$y_{t+1} = \arg\max \{P(\cdot|x, y_{0:t})\}$$

$$L(M(x)) = l \text{ s.t. } \begin{array}{ll} y_t = \textbf{eos} & \\ y_i \neq \textbf{eos} & \forall i < t \end{array}$$

	Input Space $\mathcal{X}$	Perturbation Space $\mathcal{S}(x, \delta)$
Image	$[-1, 1]^{m \times n}$	$\{x' \in \mathcal{X} \mid \|\|x' - x\|\|_\infty \leq \delta\}$
Text	V^n	$\{x' \in V^n \mid \sum_{i=1}^{n} \mathbb{1}[x_i = x'_i] \leq \delta \cdot n\}$

Our Approach

Testing: Adversarial Attacks

*Find **x'** that maximizes **L(M(x'))***

$$x' = \Pi_{\mathcal{S}(x,\delta)}(x + \alpha \nabla_x J(x))$$

Project to $\mathcal{S}(x, \delta)$ — Objective: $L(M(x))$

Projected Gradient Descent (PGD) Attack

Challenge 1: ***J(x)*** is Non-Differentiable

Greedily minimize stop probability

$$\tilde{J}(x) = \sum_{t=1}^{k} \max\left\{y_t[\textbf{eos}] - \max_{z \neq \textbf{eos}} y_t[z], \epsilon\right\}$$

Challenge 2: Discrete Inputs (Seq2Seq)

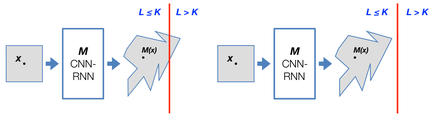
Continuous Relaxation with Gumbel Softmax

$$x = (x_1, \ldots, x_n) \rightarrow \tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$$

$$x_i = \begin{bmatrix} \cdots \\ 0 \\ 1 \\ 0 \\ \cdots \end{bmatrix}_{|V|} \xrightarrow{\text{logit}} \pi_i = \begin{bmatrix} \cdots \\ -1 \\ 1 \\ -1 \\ \cdots \end{bmatrix}_{|V|} \rightarrow \tilde{x}_i \sim \text{Gumbel}(\pi_i)$$

Verification: Constraint Encoding

*Prove ∀**x**: **L(M(x)) ≤ K***



Exhaustive Search with Mixed Integer Programming (MIP) Solver

Objective:

$$\max_{t=1\ldots K} \min_{z \neq \textbf{eos}} \left[\max_{z \neq \textbf{eos}} y_t[z] - y_t[\textbf{eos}]\right] > 0$$

ArgMax Encoding:

$$o_{\max} = \max_{y \in \mathcal{Y}}(o_y)$$

$$\Rightarrow \quad o_{\max} \geq o_y, \quad \delta_y \in \{0, 1\} \quad \forall y \in \mathcal{Y}$$

$$o_{\max} \leq o_y + (\mathbf{u} - l_y)(1 - \delta_y) \quad \forall y \in \mathcal{Y}$$

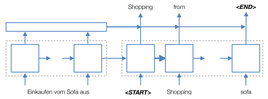
$$\sum_{y \in \mathcal{Y}} \delta_y = 1$$

Scalable but No Guarantee ⟷ *Complete but Expensive*

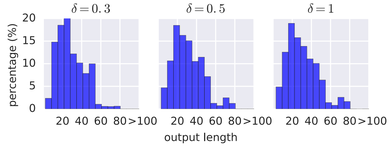
Experiments

Machine Translation (Seq2Seq)

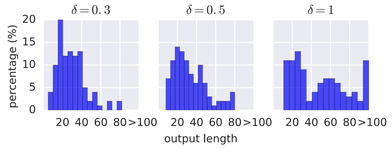
German to English
27.6 BLEU score, Vocab size: 36,548



1. Perturb test inputs randomly



2. Perturb test input with PGD attack



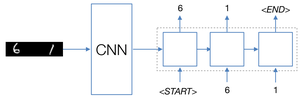
Example:

I: *Die Waffe wird ausgestellt und durch den Zaun ubergeben.*
O: *The weapon is issued and handed over by the fence.* ***eos***

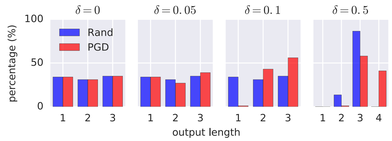
Adv Input: *Die namen name descri und ames utt origin i.e. meet grammatisch.*
Adv output: *names name names name names grammatically name names* ***eos***

Multi-MNIST (Image to text)

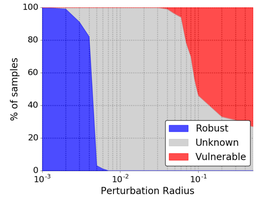
CNN — Relu-RNN
91.2% test accuracy



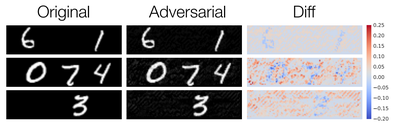
Output size distribution for perturbed inputs.



% of verified / vulnerable samples at different δ
(Verification: 30 minutes timeout for each image)



Attack: δ = 0.25



Original Output: [6, 1], [0, 7, 4], [3]
Adversarial: [6, 1, 1], [0, 1, 4, 3], [3, 3, 5, 3]