Residue-Based Natural Language Adversarial Attack Detection

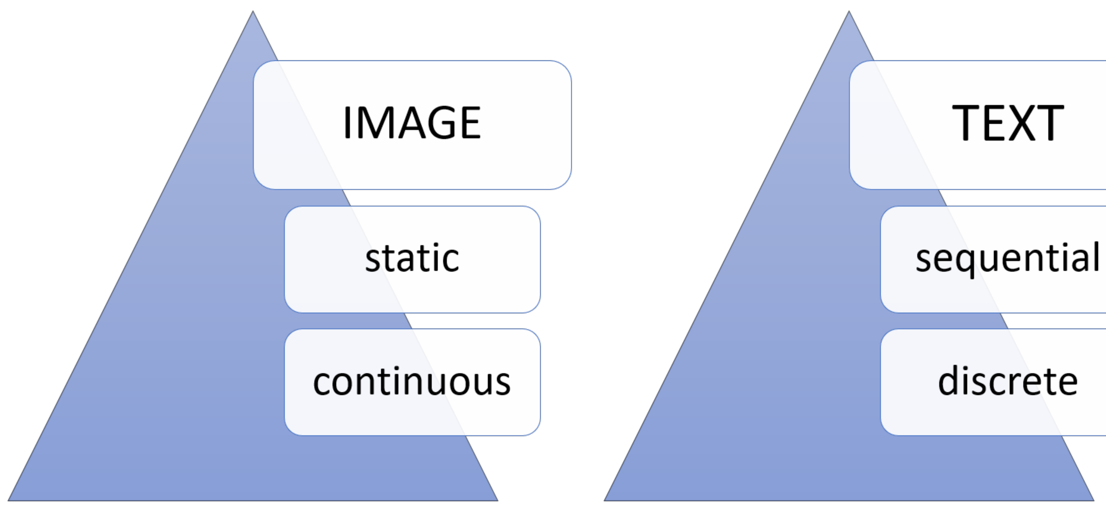
V. Raina, M.J.F. Gales {vr313,mjfg}@eng.cam.ac.uk

ALTA Institute / Department of Engineering, University of Cambridge

UNIVERSITY OF CAMBRIDGE

1. Introduction

- Deep learning systems are susceptible to adversarial attacks: small changes at an input can cause large, undesired changes at the output.
- The characteristics of inputs in different domains are very different:

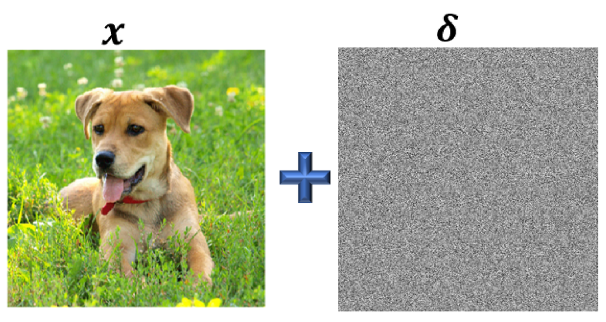


- Hence, will adversarial attack behaviour differ for image processing and natural language processing (NLP) systems?
- Should detection approaches then be tailored to the type of input?
- This work introduces a *residue-based* detection approach to specifically exploit the characteristics of inputs to NLP systems.

2. Adversarial Attacks

- A perturbation $\boldsymbol{\delta}$ at the input $\mathbf{x}$, causes a system $\mathcal{F}_{\hat{\theta}}$ to mis-classify. The perturbation has to be imperceptible.

IMAGE



$\mathcal{F}_{\hat{\theta}}(\mathbf{x}+\delta) \neq \mathcal{F}_{\hat{\theta}}(\mathbf{x})$

$||\delta||_p \leq \epsilon$

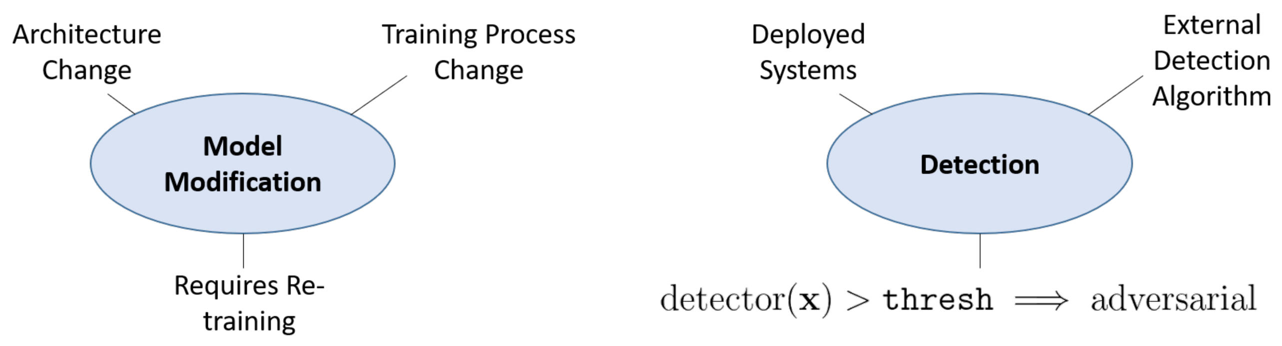
TEXT

	I enjoy sunshine	$w_{1:L}$
Synonym Substitution	I adore rays	$w'_{1:L'}$
Concatenation	I enjoy sunshine scary obese ink	$w'_{1:L'}$

$\mathcal{F}_{\hat{\theta}}(w_{1:L}) \neq \mathcal{F}_{\hat{\theta}}(w'_{1:L'})$

$\mathcal{L}_e(w_{1:L}, w'_{1:L'}) \leq N$

3. Defence

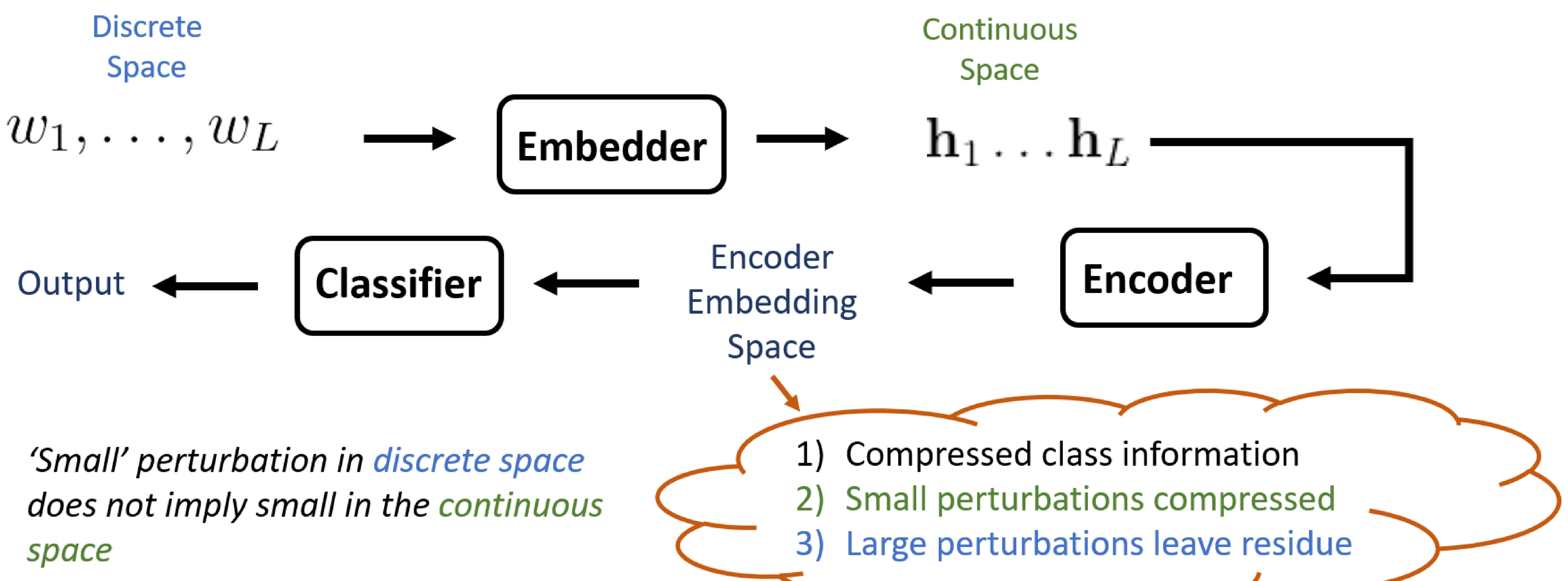


- This work focuses on detection. Methods from the image and text domain are used as baselines:

IMAGE
- Mahalanobis Distance (**MD**)
- Uncertainty (**Unc**)

TEXT
- Frequency Guided Word Substitution (**FGWS**)
- Perplexity (**Perp**)

4. Residue Detection



- We make two hypotheses:
  1. Adversarial samples in an encoder embedding space result in larger components (*residue*) in central PCA eigenvector components than original examples.
  2. The residue is only significant (detectable) for systems operating on discrete data (e.g. NLP systems).
- This motivates a simple linear classifier as an adversarial attack detector in the encoder embedding space, $\mathcal{F}_{\text{en}}(\mathbf{x})$ with parameters $\mathbf{W}$, b,

$$P(\text{adv}|\mathbf{x}) = \sigma(\mathbf{W}\mathcal{F}_{\text{en}}(\mathbf{x}) + b)$$

.

5. Experiments

- 5 different datasets (4 classification, 1 regression).
- Different Transformer-based systems trained for each task.

	# Train	# Test	# Classes	Transformer	Performance	
IMDb	25,000	25,000	2	BERT	Acc: 93.8%	Classification
twitter	16,000	2000	6	ELECTRA	Acc: 93.3%	Classification
AG NEWS	120,000	7600	4	BERT	Acc: 94.5%	Classification
DBpedia	560,000	70,000	14	ELECTRA	Acc: 99.2%	Classification
Linguaskill	900	202	1	BERT	PCC: 0.749	Regression

6. Results

- Probability Weighted Word Saliency (PWWS) used for substitution.
- A Greedy Universal approach used for concatenation attacks.
- F1 Score measures success of each detector.

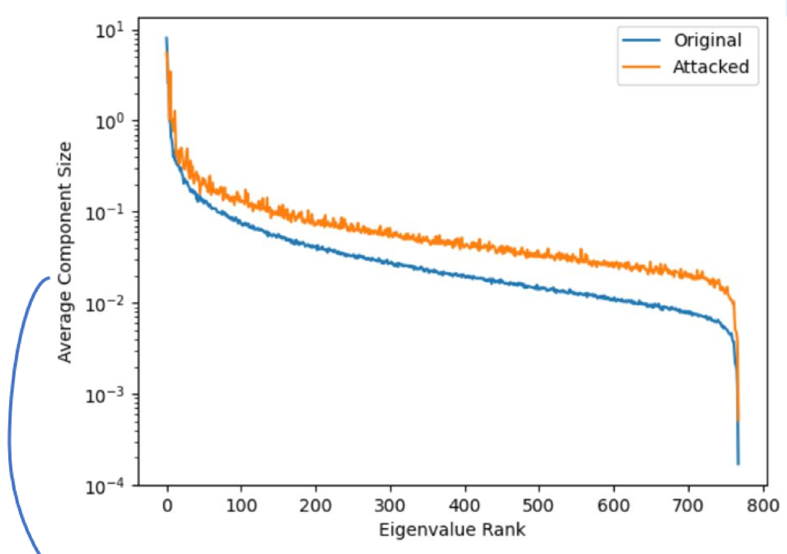
	Attack	N	Impact	Res	Perp	FGWS	MD	Unc
IMDb	sub	25	Fool: 0.70	0.91	0.68	0.87	0.67	0.75
twitter	sub	6	Fool: 0.70	0.84	0.67	0.76	0.67	0.78
AG NEWS	sub	40	Fool: 0.70	0.95	0.69	0.89	0.68	0.75
DBpedia	sub	25	Fool: 0.52	0.80	0.67	0.82	0.68	0.90
Linguaskill	con	3	Score: +0.51	0.99	0.68	0.91	n/a	0.81

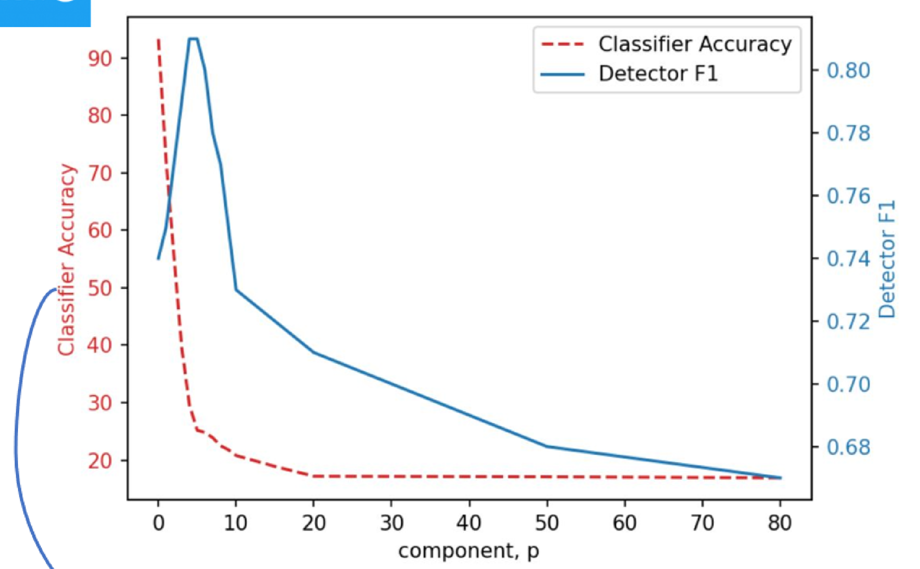
Attack: Attack, N, Impact. Defence: Res (Ours); Perp, FGWS (Text); MD, Unc (Image).

7. Analysis

- Is residue in the central PCA eigenvector components of the encoder?

twitter





$\rho_i = \frac{1}{J}\sum_{j=1}^{J} \rho_{i,j}$

$\rho_{i,j} = \left|\mathcal{F}_{\text{en}}(\mathbf{x}_j)^{\mathrm{T}}\mathbf{q}_i\right|$

Eigenvector

$\mathcal{F}_{\text{cl}}(\mathbf{x}^{(p)})$

$\mathbf{x}^{(p)} = F_{\text{en}}(x) - \sum_{i\notin[p,p+w)} \mathbf{q}_i^T F_{\text{en}}(x)\mathbf{q}_i$

$w = 5$

- Does detectable residue only exist for discrete input domains?
  - Project Gradient Descent attack for continuous space.
  - Substitution attack for discrete space.

	Domain	Attack	Res	Unc	MD
	twitter NLP-Disc	$N = 3$	0.80	0.74	0.67
		$N = 3$	0.84	0.78	0.67
Attack word embeddings	twitter NLP-Cont	$\epsilon = 0.1$	0.67	0.71	0.68
		$\epsilon = 0.3$	0.67	0.80	0.85
Attack 2 bit quantized images	CIFAR Img-Disc	$N = 200$	0.78	0.67	0.70
		$N = 400$	0.84	0.68	0.72
	CIFAR Img-Cont	$\epsilon = 12$	0.68	0.70	0.72
		$\epsilon = 48$	0.83	0.81	0.87

Attack: Domain, Attack. Defence: Res, Unc, MD.

8. Conclusions

- Adversarial attack behaviour in systems with discrete inputs (text) is different than systems with continuous inputs (images).
- Adversarial attack detection systems should be tailored to the input form.
- *Residue*-detection introduced in this work is found to be a powerful detection approach for NLP systems, where inputs are discrete and sequential.