SFU SIMON FRASER UNIVERSITY

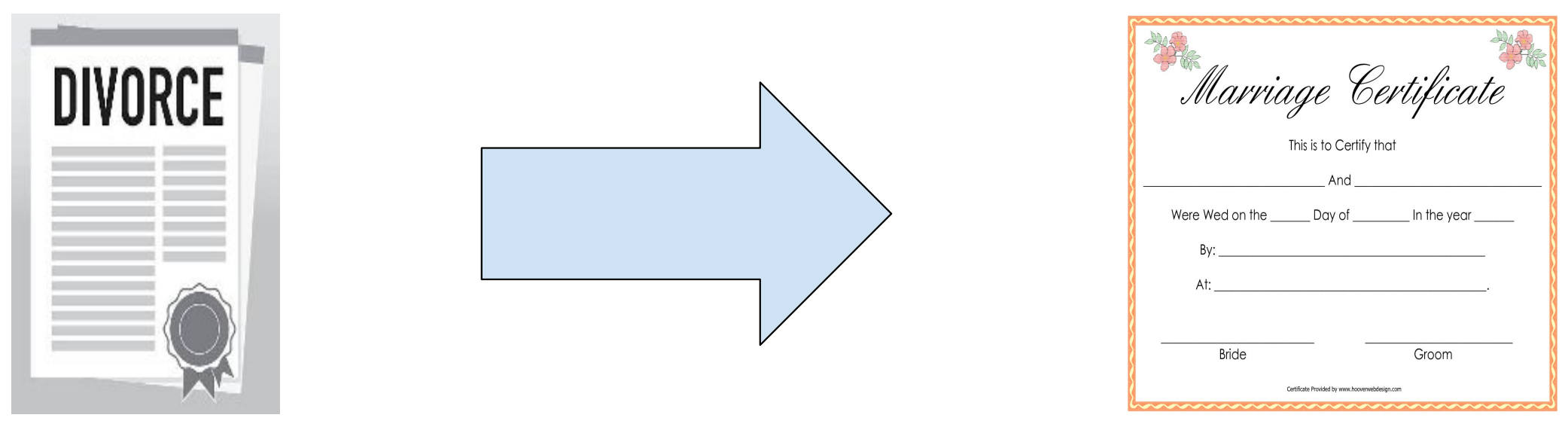
A Simple Method for Defending Against Adversarial Attacks in NLP

Xiang Xu, Kevin Joseph, Xiang Huang

CMPT 825 NLP

Introduction

Adversarial training is a common method for robust neural networks. One simple way called Adversarial Examples is to alter input data to cause mistakes in neural networks.



In this project we will focus on a single word-level adversarial attack, as well as two potential defense methods in NLP.

Attack Method

- The Probability Weighted Word Saliency (PWWS) attack ensures lexical correctness with few grammatical errors and **little semantic shifting.**
- Replaces regular words with synonyms and named entities with other entities of the same entity type.

$$\mathbf{x} = w_1 w_2 \ldots w_i \ldots w_n \qquad w_i^* = R(w_i, \mathbb{L}_i)$$

$$\mathbf{x}_i' = w_1 w_2 \ldots w_i' \ldots w_n \qquad = \arg\max_{w_i' \in \mathbb{L}_i} \{P(y_{\mathbf{true}}|\mathbf{x}) - P(y_{\mathbf{true}}|\mathbf{x}_i')\}$$

Defense Methods

Spectral Normalization

- Spectral Normalization (SN) controls the lipschitz constraints of the network. This ensures that small adversarial changes are less likely to affect the final result.

Maximizing Word Embeddings Dissimilarity

- Our hypothesis is that during an attack, synonyms close to semantically different words are found using PWWS.
- Our solution is to maximize the dissimilarity between word embeddings so that semantically different words are very far away from each other. We explicitly formulate the metric using *Hamming distance*:

$$dist_H(w_i, w_j) = \frac{1}{2}(L - w_i^T w_j)$$

- One benefit of using Hamming distance is that the larger the minimum Hamming distance is for all word embeddings in a document, the more corruptions the model can have while still being classified correctly. Formally, the output can recover from **V** bits of errors given:

$$d_{min}(Doc) = min(dist_H(x, y) : x, y \in Doc, x \neq y)$$

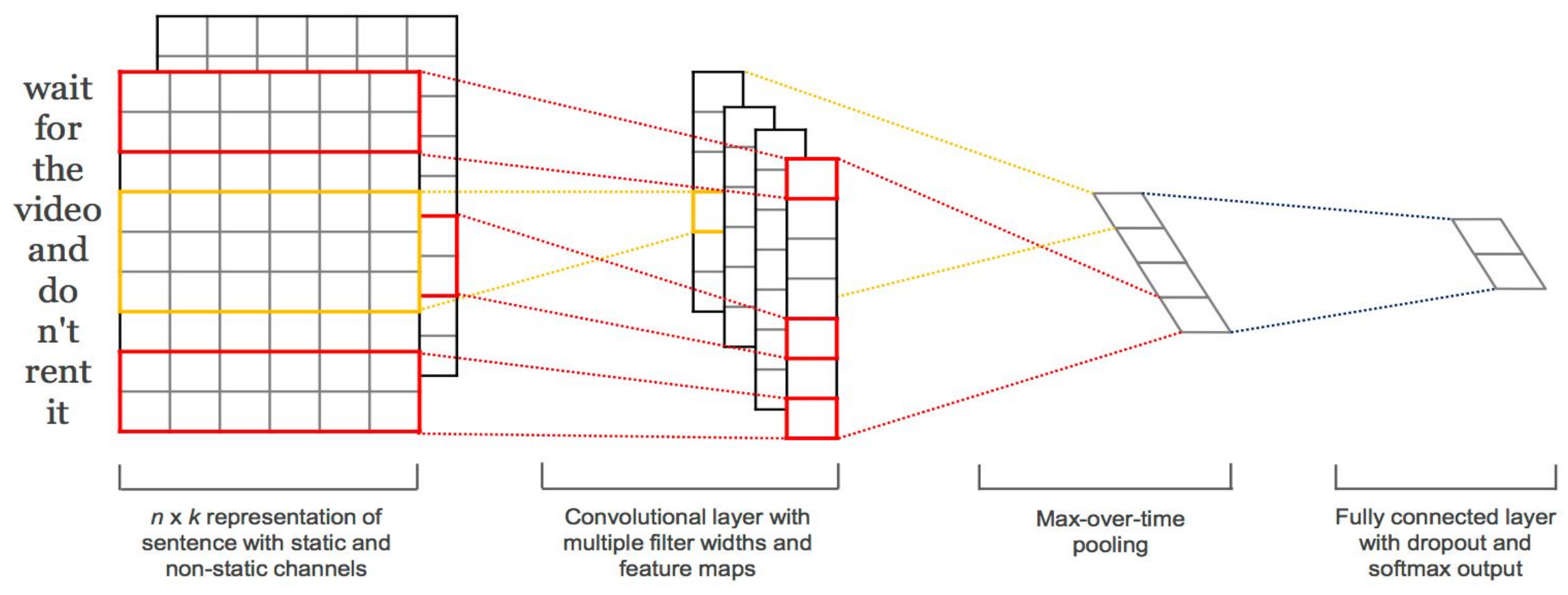
$$V \leq \lfloor \frac{d_{min}(Doc) - 1}{2} \rfloor$$

- To ensure the word embeddings are in the Hamming space, we further add a binary quantization loss.

Dataset + Model

IMDb Sentiment Analysis

- Positive / Negative Classes Only
- 25000 Train / 200 Test


wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Model and Eval Criteria

- PWWS will iterate until it either finds an adversarial attack, or none
- We evaluate using the following metrics:
  - Clean Accuracy: percentage of correctly classified clean inputs
  - Substitution rates: percentage of words or named entities substituted
  - Attack Success Rate: percentage of successful adversarial inputs
  - Avg Replacement Similarity: cosine similarity of replaced word with original word

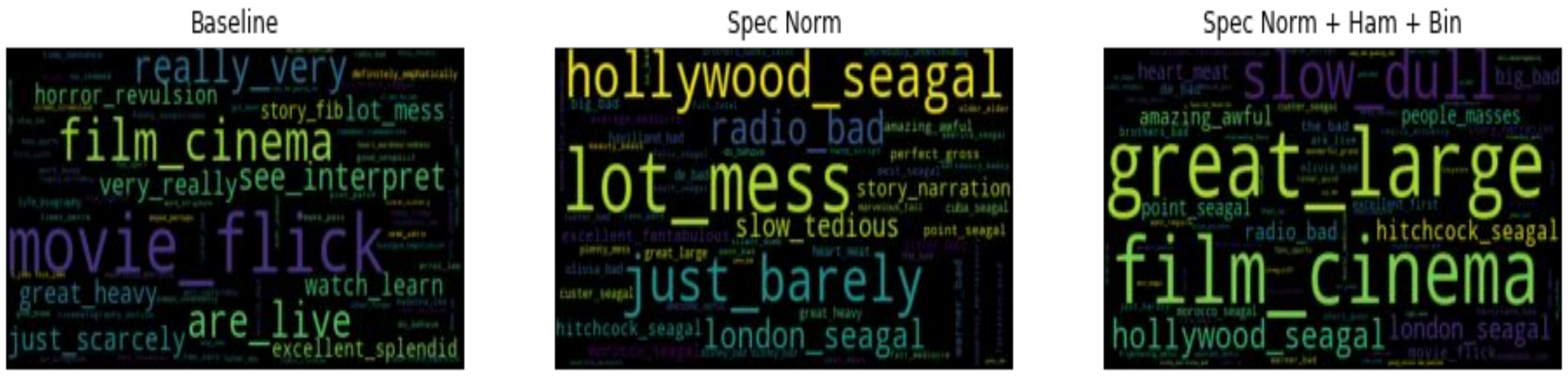
Results

Method	Clean Accuracy ↑	Attack Success Rate ↓	Word Sub Rate ↑	NE Sub Rate ↑	Avg Replacement Sim ↓
Baseline	**90.5%**	92.0%	5.10%	8.20%	0.986
SN	89.5%	85.0%	6.90%	16.10%	0.928
SN + Binary	89.0%	73.5%	10.04%	9.13%	**0.837**
SN + Binary + Cosine Similarity	89.5%	91.5%	5.63%	11.02%	0.856
SN + Binary + Hammming	88.5%	**70.5%**	**11.64%**	**11.37%**	0.846

Generated Adversarial Example

Baseline	Spectral Norm (SN)	SN + Ham + Bin
The film begins with a jaded professor haranguing his class because the students have the audacity to not be as incredibly brilliant as he is! You can tell very quickly that this man is a total cynic--finding the value in practically nothing but sticking to his own inner sense of self-importance. Additionally, he seems tired and bored with the monotony of life.Later in the film, he walks into a bank robbery and manages to annoy the robbers so much that one of them shoots him in the head. Oddly, this is only half-way through the film and what followed was a very bizarre narration of the final seconds of his life. This is when the film becomes exciting because the style of the narration is just like one of this literature professor's novels--one that is intelligently written and says things the way we wish we could all say them.See this weird film--it's **amazingly**/**astonishingly** compelling and not like anything I've ever seen before.	The film begins with a jaded professor haranguing his class because the students have the audacity to not be as incredibly **brilliant**/**bright** as he is! You can tell very quickly that this man is a total cynic--finding the value in practically nothing but sticking to his own inner sense of self-importance. Additionally, he seems tired and bored with the monotony of life.Later in the film, he walks into a bank robbery and manages to annoy the robbers so much that one of them shoots him in the head. Oddly, this is only half-way through the film and what followed was a very bizarre narration of the final seconds of his life. This is when the film becomes exciting because the style of the narration is just like one of this literature professor's novels--one that is intelligently written and says things the way we wish we could all say them.See this weird film--it's amazingly compelling and not like anything I've ever seen before.	The film begins with a jaded professor haranguing his class because the students have the audacity to not be as incredibly **brilliant**/**vivid** as he is! You can tell **very**/**really** quickly that this man is a total cynic--finding the value in practically nothing but sticking to his own **inner**/**internal** sense of self-importance. Additionally, he seems **tired**/**old** and bored with the monotony of life.Later in the film, he walks into a **bank**/**coin** robbery and manages to **annoy**/**bother** the robbers so much that one of them shoots him in the head. Oddly, this is only half-way through the film and what followed was a very bizarre narration of **the**/**minutes** **final**/**minutes** **seconds**/**minutes** of his **life**/**biography**. This is when the film becomes exciting because the style of the **narration**/**yarn** is **just**/**barely** like **one**/**zero** of this literature professor's novels--**one**/**zero** that is intelligently written and says things the way we wish we could all say them.See this weird film--it's amazingly compelling and not like anything I've ever seen before.

Replacement Word Clouds



Conclusions

- While a minor loss in accuracy is suffered on clean examples, our method is much more robust to adversarial attacks with PWWS
- The hamming word embedding modification + quantization loss also improves upon spectral normalization results