Adversarial Examples for Neural Automatic Essay Scoring Systems

Klint Kanopka (kkanopka@stanford.edu), David Lang (dnlang86@stanford.edu)

Stanford University

Abstract

Automatic essay scoring (AES) systems appear in increasingly important standardized testing applications. We investigate the vulnerability to score manipulation of a neural AES model to see if adversarial attack strategies can be developed for neural AES systems. We present a series of experiments that leverage anchors, a neural network interpretation tool, as a strategy to reduce the search space for adversarial example generation in essay-length NLP input. We find anchors can be used to identify essays with scores that are sensitive to perturbation and, given substitution into sensitive essays, anchor-based adversarial attack strategies outperform similarly constructed non-anchor-based strategies.

Research Questions

- Can adversarial attack strategies be developed for neural automatic essay scoring systems?
- To what extent are automatic essay scoring systems susceptible to score manipulation?

Dataset

- Hewlett Foundation: Automatic Essay Scoring competition on Kaggle
- Subset of the data: 1800 persuasive essays written by 10th graders
- Essay Prompt regarding censorship in libraries
- Graded (by humans) on a scale from 1-6

Approach

- Develop a neural AES system (black box attack)
- Identify anchors for specific example essays
- Identify candidate essays for perturbation
- Explore anchor-based adversarial attack strategies

Evaluation

- AES Model: Prioritize Accuracy, look to MSE as a posterior check
- Adversarial Examples: Magnitude and reliability of induced score variance

Neural AES Model

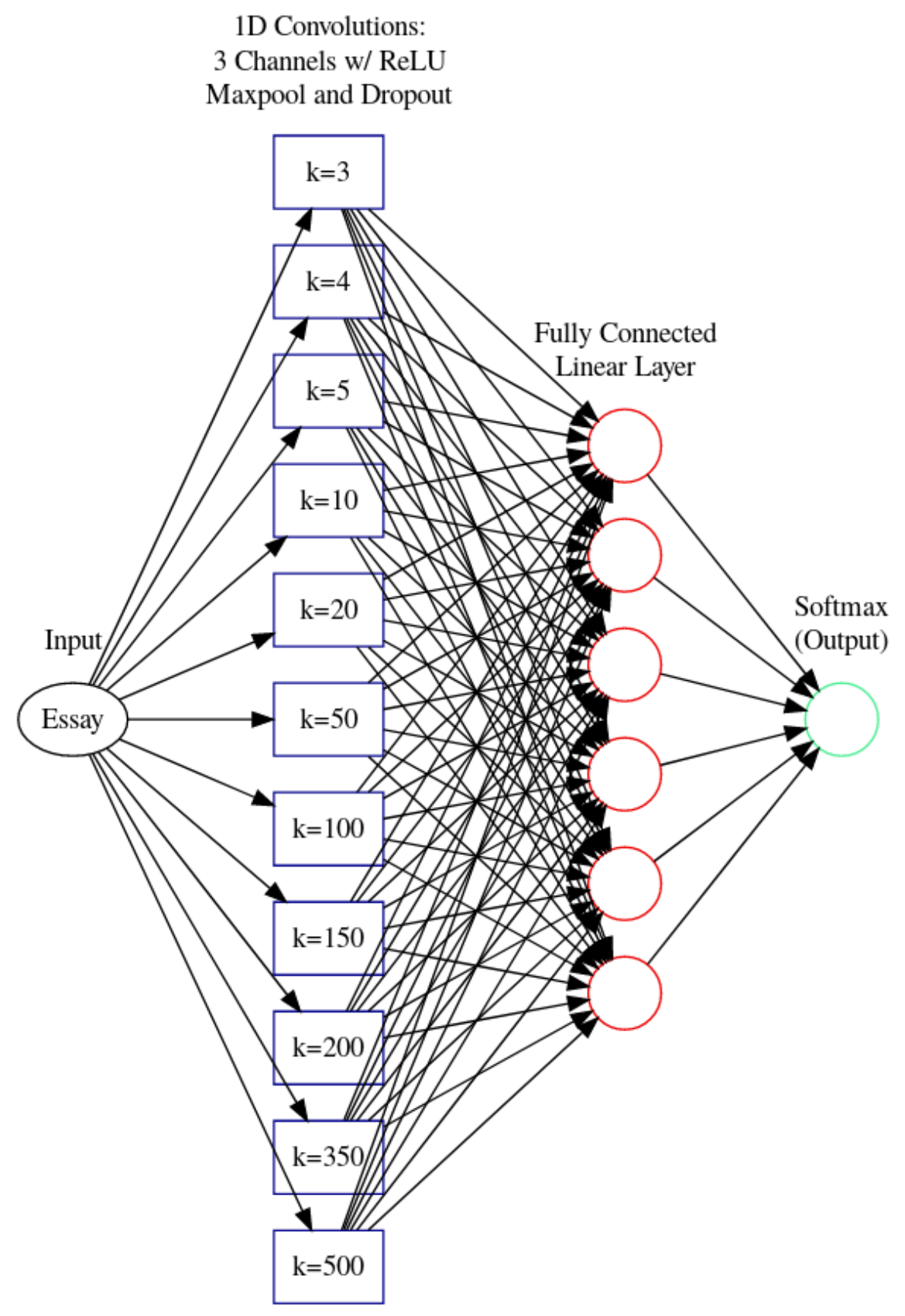

Figure 1: Neural AES System Architecture

- Input is Word2Vec embeddings
- Trained for 200 epochs, $\lambda = 0.01$
- Model Accuracy: 0.867, MSE: 0.150

Experiments

Anchor Search

- Searched for anchors by substituting *UNK* tokens for individual words - looked for minimum input to "anchor" an essay to a score
- Computationally expensive. Very slow on essay-length input.

Model Stability

- Perturbed all essays by shuffling words to investigate stability of scoring decisions under common adversarial attacks.
- Perturbed all essays by inserting/substituting/appending anchors and anchor synonyms to investigate stability of scoring decisions and identify essays that may be sensitive to manipulation.

Score Manipulation

- Using essays identified as sensitive to perturbation, systematically investigated resulting score distributions under perturbation.
- Substituted anchors, anchor synonyms, and control words for all possible words in an essay (either one substitution at a time or repeated substitutions).

Results

Example anchors identified include:

- "perhaps"
- "shelves" AND "understanding" AND "The" AND "Offended"
- "censorship" AND "Books" AND "The" AND "content"
- "freedom"

Applying anchor-based search strategies identified essays as sensitive to perturbation. Using anchor-based perturbations, scores in these essays could be manipulated.

Substitution	Mean Δ Score (Sensitive)	Mean Δ Score (Control)
Perhaps*	0.110	0.000
Freedom*	0.341	0.000
Maybe	0.030	0.000
The	0.093	0.000

Table 1: Words denoted with * are identified anchors.

Discussion

Our approach has a number of benefits:

- Model agnostic - techniques apply to any neural architecture
- Systematically provides candidate anchors to help understand scoring decisions and construct adversarial attacks.
- Systematically identifies inputs sensitive to perturbation to understand scoring decisions and select candidate essays for perturbation.

Despite this, there are drawbacks:

- The search space around essay-length input is enormous.
- Anchor search is computationally expensive.
- Perturbation effectiveness is sensitive to the location in which text is injected into the essay.

References

Ellis B Page and Nancy S Petersen. "The computer moves into essay grading: Updating the ancient test."*Phi delta kappan*, 76(7):561, 1995.
Yoon Kim."Convolutional neural networks for sentence classification."*arXiv preprintarXiv:1408.5882*, 201
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." *Thirty-Second AAAI Conference on Artificial Intelligence.* 2018.