Adversarial Techniques for Neural Automatic Essay Scoring

CSCI S-89A Deep Learning for Natural Language Processing

Lee, Michael
Nickerson, Micah

Problem Statement: With the increasing popularity of Automated Essay Scoring (AES) systems, it is critical to ensure the integrity of these systems against adversarial attacks that try to manipulate the model to achieve a desired score prediction result. In this project, our team develops a Dual Score Convolutional Neural Network (CNN) AES system and then tests the strength of this system against different types of adversarial attacks.

Data Set: This project uses an essay set from the Hewlett Foundation's Automated Student Assessment Prize (ASAP) dataset hosted on Kaggle.

Model: We built a CNN Automated Essay Scoring system based on a CNN AES scoring model implemented by Lang, et al. We further expanded our model as Dual Score Automated Essay Scoring system by predicting scores for both writing content (score range of 1-6) and grammatical accuracy (score range of 1-4).

Uses/Benefits: Automated Essay Scoring systems are used by the education industry in standardized testing environments to reduce the burden of essay grading on human readers.

Drawbacks/Challenges: Automated Essay Scoring systems do not have a high level of accuracy, human graders are still required. AES systems are also susceptible to adversarial attacks.

Results: Our Dual Score CNN Automated Essay Scoring model has an accuracy rate of 57% for estimating both the writing content score and the language conventions score correctly. Our model can perform better than benchmark professional AES systems but not as well as current research AES examples.

The adversarial attacks we implemented were not effective in achieving a higher AES score prediction. Instead, we found that the adversarial attacks had a slight negative impact on AES score predictions.

Project Files
Link to Dataset: https://www.kaggle.com/c/asap-aes
Link to Model: https://github.com/mjnickerson/csci-89a-final_project
Link to Presentation Video: https://youtu.be/V0C0RNsNceY

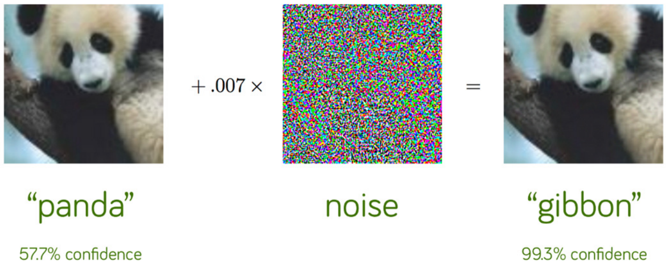
Introduction

Neural automated essay scoring (AES) is a supervised learning predictive neural network to "read" a specific essay set and classify a score on an assigned point scale. They are predominantly implemented in the educational testing industry for standardized testing applications; the goal is to lessen the burden on human readers by pre-scoring essays and flagging those essays which require manual verification.

However, there are concerns that the automated essay scoring systems can be tricked or manipulated to yield a higher score than deserved. A classic feature-based AES model can be tricked by single-word replacement attacks or nonsense attacks (Lang 2019). Even neural AES systems are vulnerable to adversarial attacks.

Adversarial attacks of a neural network use a carefully chosen input to trick the neural network the network with a change imperceptible to a human observer to yield a different outcome (Gilmer 2017). An example of an adversarial attack on an image classification neural network is shown below in **Figure 1**. We see that the neural network classifies the first image as a panda with 57.7% confidence. However, after adding some noise which is imperceptible to a human, the neural network now classifies the image as a gibbon with 99.3% confidence.

Figure 1 - Example of Adversarial attack, per Jain (2019).



There are a couple of purposes for developing and implementing adversarial attacks on neural networks. One motivation is to trick the neural network for nefarious purposes, such as to improperly receive a higher score on an AES scored essay than deserved. On the other hand, developers of AES systems also want to develop adversarial attacks to test their neural network systems and to identify vulnerabilities. Jain outlines two forms of defenses against adversarial attacks: adversarial training and defensive distillation. In adversarial training, a number of adversarial attacks are generated and then the neural network is trained on how to avoid these attacks. In defensive distillation, a secondary model is trained with a surface which is smoothed, making it difficult for an attacker to discover adversarial input tweaks.

Kanopka and Lang developed a methodology of testing of adversarial attacks on Natural Language Processing applications, such as AES systems by identifying anchors, which represent the minimum input to a model required to ensure a desired score prediction (Lang 2019). Once anchors in a neural network model are identified, then the essays can be manipulated using different word perturbations. Some of the word perturbations include:

- Shuffling - the words in an essay are randomly shuffled
- Appending - the target anchor word is appended to the end of the essay with no additional modifications
- Insertion - the target anchor word is randomly inserted into the essay with no additional modifications
- Substitution - a word in the essay is replaced with the target anchor word with no additional modifications

Kanopka and Lang further explored the use of the following substitution based adversarial attack techniques for manipulating the AES score predictions:

- Progressive overload - words in the essay are cumulatively replaced with the target anchor word (ex. The first word in the essay replaced by the target anchor word for the first example, first two words in the essay replaced by the target anchor word for the second example, etc.). The objective of progressive overload is to determine how the model is affected by the increasing and repetitive use of an anchor word.
- Single Substitution - a word in the essay are replaced with the target anchor word (ex. The first word in the essay is replaced by the target anchor word for the first example, the second word in the essay is replaced by the target anchor word for the second example, etc.). The objective of single substitution is to test the model's sensitivity to the location of a word perturbation.

Methodology

Our team's objective is to explore the adversarial techniques listed above and understand how effective the Automated Essay Scoring systems are against these adversarial attacks. Our approach follows the methodology used in Kanopka and Lang's study: first building an automated essay scoring system and then developing adversarial attacks against the AES.

In our project, we first build a CNN-based Automated Essay Scoring system. This AES scores student written essays based on an essay prompt. After developing the Automated Essay Scoring system, we implement the adversarial attack strategies that Kanopka and Lang demonstrated and we evaluate if there are changes in the predicted score output due to the adversarial attacks. For the selection of anchor words which influence the model score, we propose to use the words "library" and "censorship" as candidate anchor words since the essay prompt asked about the student's views of censorship in libraries. As a control, we also use the word "the" in the adversarial attacks to see if anchor words have a larger change in AES score prediction.

We first conduct the adversarial attacks on the essays using shuffling, appending, and insertion word perturbations and then we try the progressive overload and single substitution attacks. After implementing these adversarial attacks on a sample of essays in the test data set, we observe the AES score predictions to see if they are different than the original, unmodified essays and determine if the adversarial attacks influence the AES score prediction.

Data Set

The dataset used for our Automated Essay Scoring system is the Hewlett Foundation's Automated Student Assessment Prize (ASAP) dataset which was made available in a Kaggle competition for developing an automated essay scoring system. This dataset is a compilation of eight datasets of essays written by seventh to tenth graders in response to different essay prompts. To match Kanopka and Lang's study, we chose to use Essay Set #2, which asks students to discuss the role of censorship in libraries. There were 2,400 student essays in this dataset and the essays were scored by human graders from a range of 1-6 for writing content and a range of 1-4 for language convention (grammatical accuracy). From exploratory data analysis, the average essay is approximately 400 words long, and has dual scores of 3 or 4.

Installation and Configuration

To generate a closed "black box" as a host for adversarial testing, we used an AES CNN system implemented by Lang, et al. in 2019 as basis of design. We also based our model on an initial sentence classification CNN created by Kim, et al in 2014. Whereas in the Lang model only half of the essay evaluation was performed (single score) our model was expanded to perform the complete consider language construction and grammar, in addition to meaning.

Our final model took the form of:

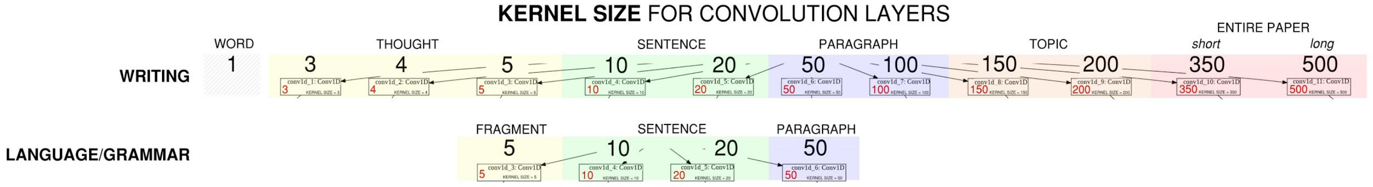
1. Each complete essay is tokenized into word tokens, and the words.are assembled into a dictionary (including both words and punctuation), with a maximum of 20,000 words.
2. Using Stanford's GloVe 300 dimension word embeddings, the words for each essay are embedded into vectors, and those embeddings are concatenated into a single vector.
3. The essay vector is then passed into two parallel CNNs (see **Figure 3**), one to score writing style (content and voice), and the other to score language conventions (grammar and punctuation).

4. In the "Domain 1" (Writing Style) Deep Convolutional Neural Network:
    a. The essay is then passed through an embedding layer, with GloVe embeddings.
    b. Eleven parallel Convolutional layer sets of 3 feature maps each convolve over the essay, using relu activation, and each with a different kernel size from the range of:
        i. 3,4,5,10,20,50,100,150,200,350,500 (see **Figure 2** below). The intent is for each colvolving over the set to identify language features at different scales, from sentence, paragraph and up to the entire paper.
    c. Learning representation of each Conv1D layer is then reduced by 50% via a Maxpooling layer.
    d. Dropout of 50% is applied to address overfitting.
    e. Essay representation is then flattened.
    f. Essay vector is passed into a fully connected dense layer of 6 neurons.
    g. Softmax Activation generates an output score, from 1-6, for Writing Style.

5. In the “Domain 2” (Grammar and Language) Deep Convolutional Neural Network:
    a. The essay is then passed through an embedding layer, with GloVe embeddings.
    b. Four parallel Convolutional layer sets (see **Figure 4**) of 3 feature maps each convolve over the essay, each with a different kernel size from the range of:
        i. 5,10,20,50 (see **Figure 2** below). The intent is for each colvolving over the set to identify language features for short sentences, long sentences, and paragraphs.
    c. Learning representation of each Conv1D layer is then reduced by 50% via a Maxpooling layer.
    d. Dropout of 50% is applied to address overfitting.
    e. Essay representation is then flattened.
    f. Essay vector is passed into a fully connected dense layer of 4 neurons.
    g. Softmax Activation generates an output score, from 1-4, for Language Conventions.

6. Both Domain 1 and Domain 2 scoring outputs are fed into a scoring engine. First performance is evaluated for both different scoring methods. Then all subsets of essay scoring accuracy are computed and their intersection recorded.
7. A confusion matrix is generated for final score, for essays scored correctly by both domains, either domain, or neither.
8. A Final Score is evaluated and produced. Only for essays that received a correct score on both domains are considered accurately scored.

Figure 2 - “Kernel Sizes - CNN layers



Our network was trained in ranges from 2 to 200 epochs, and found that an epoch count of 8 was highest accuracy for both Domain 1 and Domain 2 scores, reaching top accuracy indicated in **Figure 5** below. These two models were packaged into two “black box” h5 models (D1_76_BLACKBOX_CNN.h5 & D2_69_BLACKBOX_CNN.h5), and isolated from the model training as a foreign entity.

At each evaluation of test essays or adversarial essays, these two model files were completely reloaded each time- the intent being as though the models were furnished by a third party as a lock to crack.

Figure 3 - "Black box" Architecture- Automated Essay Scoring CNN

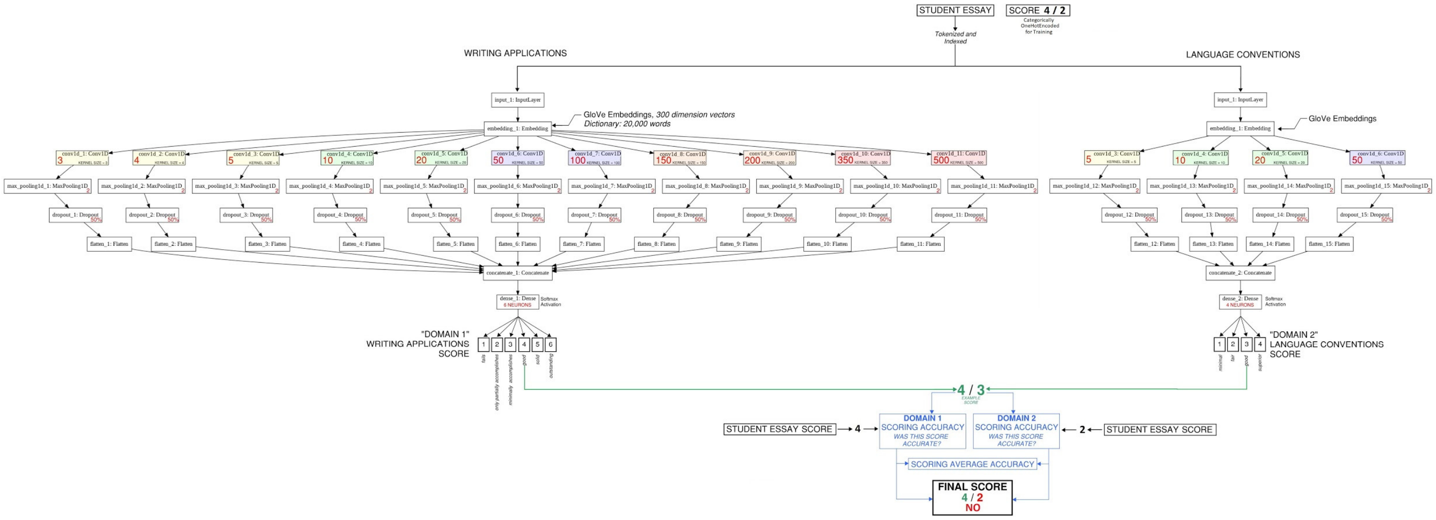
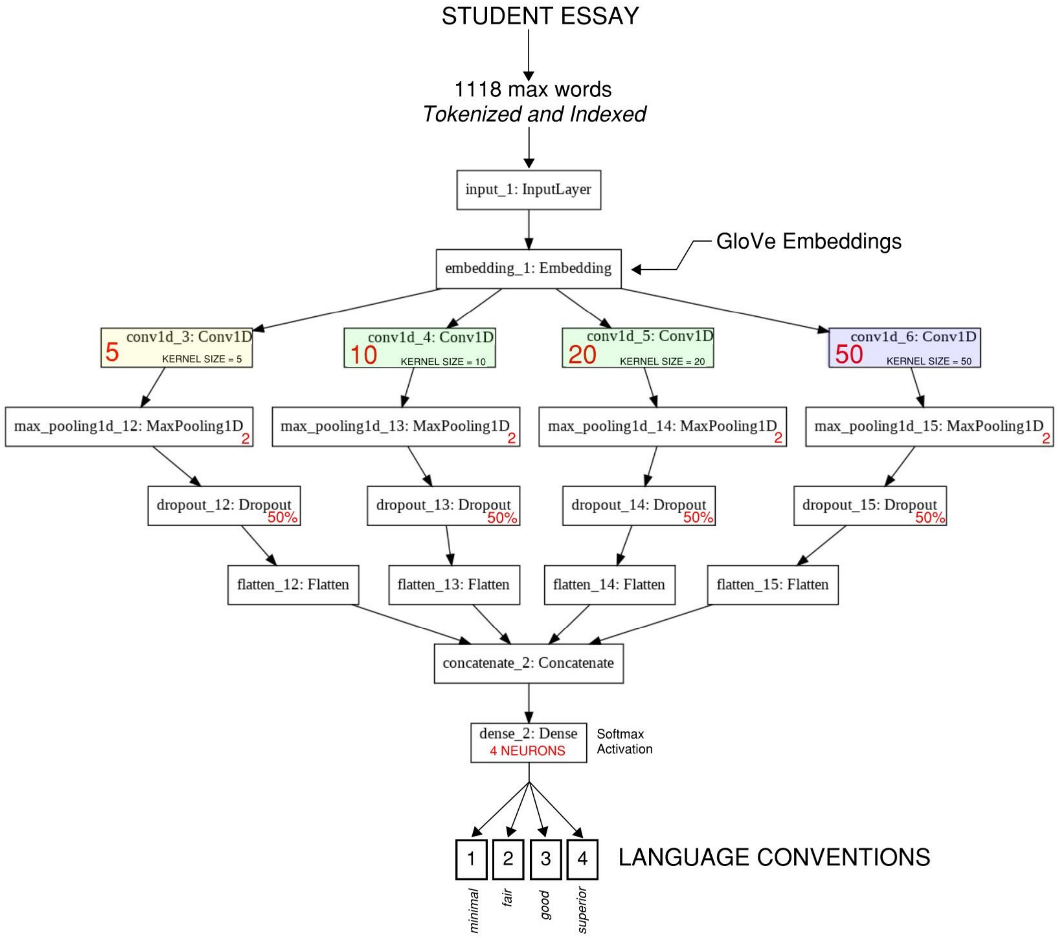
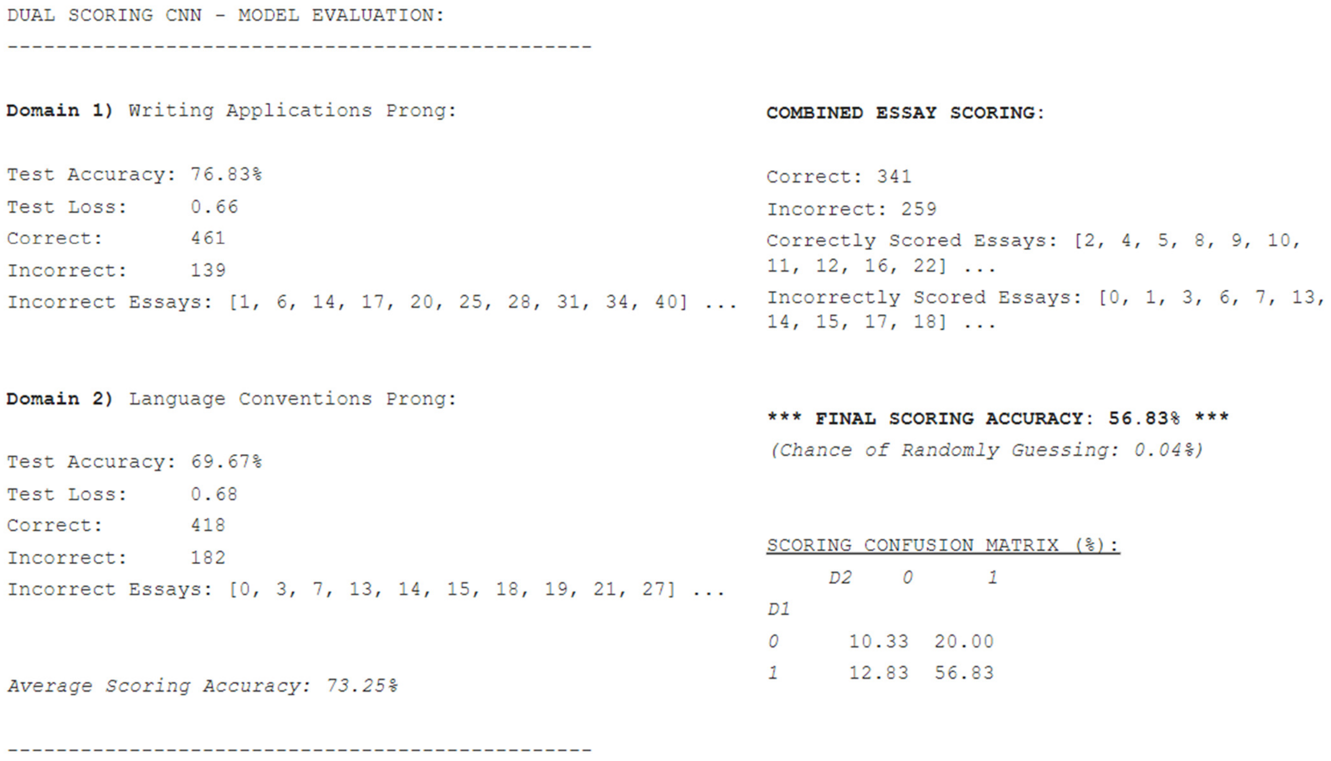


Figure 4 - Enlargement - "Domain 2" Portion - AES CNN



Automated Essay Scoring “Black Box” - Test Results

Figure 5 - “Black box” Model Performance - Study Automated Essay Scoring CNN



```
DUAL SCORING CNN - MODEL EVALUATION:
------------------------------------------------

Domain 1) Writing Applications Prong:

Test Accuracy: 76.83%
Test Loss:     0.66
Correct:       461
Incorrect:     139
Incorrect Essays: [1, 6, 14, 17, 20, 25, 28, 31, 34, 40] ...


Domain 2) Language Conventions Prong:

Test Accuracy: 69.67%
Test Loss:     0.68
Correct:       418
Incorrect:     182
Incorrect Essays: [0, 3, 7, 13, 14, 15, 18, 19, 21, 27] ...


Average Scoring Accuracy: 73.25%

------------------------------------------------
```

```
COMBINED ESSAY SCORING:

Correct: 341
Incorrect: 259
Correctly Scored Essays: [2, 4, 5, 8, 9, 10,
11, 12, 16, 22] ...
Incorrectly Scored Essays: [0, 1, 3, 6, 7, 13,
14, 15, 17, 18] ...


*** FINAL SCORING ACCURACY: 56.83% ***
(Chance of Randomly Guessing: 0.04%)


SCORING CONFUSION MATRIX (%):
     D2     0      1
D1
0      10.33  20.00
1      12.83  56.83
```

To benchmark performance of the AES built, we selected three references - the high score of ASAP student competition from 2013, the most recent AES manual published by ETS (for the GRE), and the chances of randomly guessing a correct set from a most likely set of scores (3 or 4).

- ASAP student competition - High Score - 81.40%
- **Our Dual CNN AES - 56.83%**
- ETS GRE Scoring System - 52.00%
- Random Guess (set of 3 and 4) - 25.00%

It is worth noting that the ASAP competition high score was for a single (Domain 1) score evaluation, so our black box accuracy is 76.83% vs 81.40% in a level competition.

Adversarial Testing Results

In Tables 1 and 2, we evaluate the change in the predicted writing content score (Table 1) and predicted language convention score (Table 2) due to the adversarial attacks. This includes adversarial attacks on both anchor and control words.

Table 1: Change in Writing Content Score due to Adversarial Attack

Attack	Score Decreased	Score Increased	No change
Shuffling	4.0%	8.0%	88.0%
Appending	2.0%	2.0%	96.0%
Insertion	2.0%	0.7%	97.3%
Progressive Overload	2.0%	0%	98.0%
Single Substitution	2.7%	0.7%	96.7%

Table 2: Change in Language Convention Score due to Adversarial Attack

Attack	Score Decreased	Score Increased	No Change
Shuffling	14.0%	14.0%	72.0%
Appending	10.0%	4.0%	86.0%
Insertion	7.3%	6.0%	86.7%
Progressive Overload	10.0%	3.3%	86.7%
Single Substitution	8.0%	2.0%	90.0%

These results suggest that in the majority of the examples, the adversarial attacks do not affect the predicted writing content and language convention scores. However, in some cases, the adversarial attack results in a score prediction higher than the original score prediction and in other cases, the adversarial attack results in a score prediction lower than the original score prediction. In each case, the change in predicted score is one point. Table 2 suggests that the adversarial attacks have a larger impact on the language convention score than the writing content score.

In Tables 3 and 4, we look at the mean score change for the predicted writing content score (Table 3) and predicted language convention score (Table 4), and compare the results for both the anchor words (“library”, “censorship”) and the non-anchor control word (“the”).

Table 3: Mean Change in Writing Content Score Prediction Due to Adversarial Attack

Attack	Anchor	Non-Anchor
Shuffling	4.0%	
Appending	0%	0%
Insertion	**-2%**	0%
Progressive Overload	**-2%**	**-2%**
Single Substitution	**-2%**	**-2%**

Table 4: Mean Change in Language Score Prediction Due to Adversarial Attack

Attack	Anchor	Non-Anchor
Shuffling	0%	
Appending	**-6%**	**-6%**
Insertion	0%	**-4%**
Progressive Overload	**-8%**	**-4%**
Single Substitution	**-7%**	**-4%**

Based on these results, we conclude that in general the adversarial attacks resulted in a lower essay score prediction. While this is evident for both the writing content and language convention scores, we observe that the magnitude of the score decrease is higher for the The only adversarial attack where we see a net increase in writing content score prediction is the shuffling attack.

Based on the progressive overload and single substitution attacks, we observed that the words which we felt were suitable anchor candidates had a lower language convention score than the control word. For the insertion attack, we see the opposite trend: the control word had a lower mean language convention score as compared to the anchor candidates.

Conclusion

In this project, we built a Dual Score Convolutional Neural Network for Automated Essay Scoring. This model performed better on predicting the writing quality than on grammatical accuracy. While our AES evaluated essays than the benchmark professional AES systems, our combined dual score accuracy of 57% is lower than current research AES examples and requires more developmental work to improve the accuracy.

The results of this study showed that adversarial attacks appeared to have a slightly negative impact on our AES score predictions. The words we thought as suitable anchor candidates did not perform any better than the control words in positively influence the AES scores, in fact the anchor candidates performed worse than the control word. Overall, these results demonstrate that the use of adversarial attacks was not effective in their intended purpose of achieving higher AES score predictions.

The success of shuffling technique indicates a vulnerability in a black box system using convolutional layers to identify features within concatenated embedded vectors. Mainly, the sequence of the words can be different while the resultant learned value is the same. This "sum of bag of words" phenomenon was expected to be resolved with feature maps spanning the entire essay length, but these feature maps were clearly down weighted (non-predictive) by the trained network. Preventing shuffling requires an additional mechanism to study and verify the sentence construction for the essay, and down score/red flag essays for manual review.

As a way to improve both AES prediction accuracy and impenetrability against adversarial attacks, the team has identified the replacement of the Convolutional Neural Network with a GRU based Recurrent Neural Network. Our initial tests of a GRU based RNN AES system performed worse than our CNN AES system, with a score prediction accuracy of 43% for Domain 1. Even with attention mechanism added, the GRU was unable to read across an entire essay of words and produce a meaningful learning representation. We found there was insufficient time to optimize this model, and chose to focus on the CNN instead. With additional research and development we envision that the GRU based RNN AES system can outperform our AES model.

Finally, further testing with an expanded or transformed dataset is recommended, as our deep convolutional neural network was rapidly overfitting on training data. An increased sample set from the 10th grade level on additional essay prompts and an identical scoring rubric would benefit training and testing.

References

Attali, J.B. a. Y. a., Automated essay scoring with e-rater v. 2.0. ETS Research. *ETS Research Report Series* **2004(2):i–21**, (2005).

Briscoe, Y. F. a. Y. a., Neural Automated Essay Scoring and Coherence Modeling for AdversariallyCrafted Input. *CoRR* **abs/1804.06898**, (2018).

Dery, N.H. a. D. a., Neural Networks for Automated Essay Grading. *Stanford University* **cs224d/huyenn**, (2016).

Gilmer, T. B. B. a. M. a. R. a. A. a., Adversarial Patch. *CoRR* **abs/1712.09665**, (2017).

Guestrin, S. S. a. M. T. a, Anchors: High precision model-agnostic explanations. *Thirty-Second AAI Conference on Artificial Intelligence*, (2018).

Jain, Anant. “Breaking Neural Networks with Adversarial Attacks.” Towards Data Science, Medium, 9 Feb. 2019, towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa.

Kaggle. ”Develop an automated scoring algorithm for student-written essays.” (2012). https://www.kaggle.com/c/asap-aes

Kim, Yoon. Convolutional neural networks for sentence classification. *arXiv preprint* **arXiv:1408.5882**, (2014).

Lang, K. K. a., Adversarial Examples for Neural Automatic Essay Scoring Systems. *Stanford University* **cs224n/15720509**, (2019).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Xiong, Z. W. a. W. a., Automated Essay Scoring System. *University of Illinois at Urbana-Champaign* **CS410/aess-project**, (2018).

Wikipedia contributors. (2019, July 1). Automated essay scoring. In *Wikipedia, The Free Encyclopedia*. Retrieved 21:14, July 29, 2019, from https://en.wikipedia.org/w/index.php?title=Automated_essay_scoring&oldid=904341402

Zhao, J. L. a. X. a., Automated Essay Scoring based on Two-Stage Learning. *CoRR* **abs/1901.07744**, (2019).