INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Universal Adversarial Attacks On Spoken Language Assessment Systems

Vyas Raina, Mark J. F. Gales, Kate Knill

ALTA Institute/ Engineering Department, Cambridge University, UK
{vr313,mjfg,kmk1001}@cam.ac.uk

Abstract

There is an increasing demand for automated spoken language assessment (SLA) systems, partly driven by the performance improvements that have come from deep learning based approaches. One aspect of deep learning systems is that they do not require expert derived features, operating directly on the original signal such as a speech recognition (ASR) transcript. This, however, increases their potential susceptibility to adversarial attacks as a form of candidate malpractice. In this paper the sensitivity of SLA systems to a universal black-box attack on the ASR text output is explored. The aim is to obtain a single, universal phrase to maximally increase any candidate's score. Four approaches to detect such adversarial attacks are also described. All the systems, and associated detection approaches, are evaluated on a free (spontaneous) speaking section from a Business English test. It is shown that on deep learning based SLA systems the average candidate score can be increased by almost one grade level using a single six word phrase appended to the end of the response hypothesis. Although these large gains can be obtained, they can be easily detected based on detection shifts from the scores of a "traditional" Gaussian Process based grader.

Index Terms: spoken language assessment, adversarial attacks, assessment malpractice

1. Introduction

With the increasing demand for English language learning, there has been a growth in popularity of automated spoken language assessment (SLA) systems. Beyond assessing a candidate's English speaking ability, it is necessary to ensure that a system is robust to malpractice. The integrity and reliability of an exam comes under threat when a candidate can take actions that result in a score that is inconsistent with the exam assessment criteria. This work explores a particular form of malpractice: adversarial attacks. Here small perturbations in the input yield significant, undesired, changes in the output. Due to the success of deep-learning (neural) systems in speech [1, 2] and natural language processing [3, 4] tasks, there is interest in evaluating the susceptibility of neural assessment systems to adversarial attacks, and how these attacks can be detected.

When an adversary has access to the the internal structure of a system, the form of adversarial attack is termed a *White Box* attack [5]. However, it is unlikely that an adversary attacking an automated spoken language assessment system will have access to the internal workings of the system. Hence, in this work only black-box, adversarial attacks are considered, where the adversary has no knowledge of the system. Black-box attacks are grouped into query based approaches [6, 7] and transfer based approaches [8, 9]. If a large number of queries is required for a successful attack, the former approach is easy to detect. Alternatively transfer-based approaches rely on similar models being susceptible to the same adversarial samples [10]. Recent studies [11, 12, 13] have demonstrated successful transfer of attacks, but only in situations where the networks are extremely similar in structure and parameters.

For spoken language assessment systems it is possible to attack either the audio signal or the word sequence uttered. As features, either expert or deep-learning based, derived from the word-sequence are found to accurately predict the grade, this work focuses on text-based attacks. A wide range of simple techniques [14, 15, 16] can be employed to construct adversarial attacks. However, due to the discrete nature of the input, the text sequence, gradient based adversarial attacks are difficult to implement [17]. A range of text based attacks and detection approaches have been described in the literature [18, 19]. For SLA systems this text is derived from a speech recognition system; thus the vocabulary is fixed. This means that attacks such as character-level replacement [20, 21] cannot be used. In this work a greedy discrete search method for the adversarial attack is adopted. In particular a universal attack is considered [22], where a single phrase is found that, for SLA, will increase the predicted score. Using a universal attack reduces the opportunity for detection, as the attack needs to only be trained once and just requires the candidate to learn a set phrase. There are a range of general approaches to adversarial attack detection [23, 24]. This paper examines the use of perplexity scores [17] and deep ensembles [25, 24] approaches, as well as a SLA specific off-topic response detection approach [26]. Additionally, a detection approach based on a second, feature-based, SLA system is also described.

This paper considers adversarial attacks on SLA systems for multi-level prompt-response free speaking tests i.e. candidates from a range of proficiency levels provide open responses to prompted questions. Based on this audio input the assessment systems must predict a score of 0-6 corresponding to the 6 CEFR [27] grades. Both feature-based assessment [28, 29, 30] and deep neural assessment approaches can be used for SLA. Though the focus of this work is neural assessment, adversarial attacks on feature-based approaches are also examined.

2. Text Adversarial Attacks

The general form of a targeted adversarial attack is

$$\hat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} \{\mathcal{F}(\boldsymbol{x}+\boldsymbol{\delta}) = t\} \quad \text{s.t. } \mathcal{H}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) < \epsilon \quad (1)$$

where t is the required target outcome from the classifier $\mathcal{F}()$, $\boldsymbol{x}$ is the observation to be attacked, $\mathcal{H}()$ is some "distance" between the observation, and ϵ is a threshold at which value the perturbation on the observation $\boldsymbol{\delta}$ is deemed to be noticeable.

This work considers adversarially attacking a free-speaking spoken language assessment system. In common with other

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Thanks to Cambridge English Language Assessment for support and access to the Linguaskill data. The authors would also like to thank members of the ALTA Speech Team.

http://dx.doi.org/10.21437/Interspeech.2020-1890

systems, assessment is treated as a regression task predicting a continuous score that is then mapped to one of the CEFR levels. The task is to maximally increase the predicted score given the ASR output [1]. Initially consider appending a fixed phrase to the end of a valid response, $\boldsymbol{w}_{1:n}$, to a prompt. Thus

$$\hat{\boldsymbol{w}}_{1:n+k} = \boldsymbol{w}_{1:n} \oplus \boldsymbol{\delta}^{(k)} = w_1, \ldots, w_n, \tilde{w}_1, \ldots, \tilde{w}_k$$

where the k word adversarial attack is $\tilde{w}_1, \ldots, \tilde{w}_k$. The cost function to be optimised can then be written as

$$\hat{\boldsymbol{\delta}}^{(k)} = \underset{\boldsymbol{\delta} \in \mathcal{V}^k}{\arg\max} \left\{ \sum_{s=1}^{S} \mathcal{F}(\boldsymbol{w}^{(s)} \oplus \boldsymbol{\delta}; \boldsymbol{\theta}) \right\} \tag{2}$$

where $\boldsymbol{\theta}$ represents the trained model parameters, $\boldsymbol{w}^{(s)}$ is the valid response for candidate s and $\mathcal{V}^k$ is the set of all k length word-sequences that can be constructed using the ASR vocabulary $\mathcal{V}$. Here, a single adversarial phrase of length k is to be used for all candidates. Although an adversarial attack could in theory be generated for each candidate's recognised word sequence, this is highly challenging for a practical system and not considered further.

As black-box adversarial attacks are most realistic for SLA systems, it is not possible to optimise the attack using knowledge of the network architecture. In this work an explicit, discrete optimisation approach is adopted. This is challenging as searching all possible words in the vocabulary is expensive requiring a large number of queries. Additionally if context-dependent word embeddings, such as BERT [31], are used then adding any word alters the embeddings for all other words. To address this problem a two stage approach is used. Initially a transfer-based approach is adopted, where a simple context-independent word-embedding based substitute system [8] is used to select a subset of words, in this case 100, from the complete vocabulary [2]. This subset can then be used to query the real system to select the optimal word. This approach is felt to be realistic as only a single universal phrase is needed for all speakers. The adversarial attack is generated in a "greedy" fashion where

$$\hat{\boldsymbol{\delta}}^{(k)} = \underset{\delta \in \tilde{\mathcal{V}}}{\arg\max} \left\{ \sum_{s=1}^{S} \mathcal{F}(\boldsymbol{w}^{(s)} \oplus \hat{\boldsymbol{\delta}}^{(k-1)} \oplus \delta; \boldsymbol{\theta}) \right\} \tag{3}$$

where $\tilde{\mathcal{V}}$ is the subset vocabulary determined by the initial system. The number of system queries thus increases linearly with the length of the adversarial attack, and the size of the subset vocabulary. The attack defined has only considered appending a phrase at the end of an utterance. In practice this is the simplest for a candidate to append to a standard response, but other positions can be considered.

Having obtained a universal phrase to attack the system it is possible to examine approaches to detect these attack phrases, which can also be incorporated into adversarial attack generation if the defence mechanism is known. The form of adversarial attack in equation (2) imposes no constraints on the words being appended to the original sequence. It is therefore possible that by using a language model of "standard" non-native speakers of English it is possible to detect the adversarial attacks [17]. In general, it is not possible to know the location of the attack. Thus it is necessary to consider the average perplexity of the complete sequence, normalised by the sequence length. As a refinement to this basic model, found in initial experiments to yield a small improvement, a grade-dependent language model was used, based on the predicted grade from the neural assessment system. The metric used to assess whether an adversarial attack is being used is:

$$\log\left(P(\hat{\boldsymbol{w}}_{1:n+k}|\hat{g})\right)/(n+k) > \beta \tag{4}$$

where an initial sentence start symbol is added as w_0, $\hat{g}$ is the predicted grade from the neural assessment system [3].

One general approach for detecting adversarial attacks is to examine the consistency of the ensemble of predictors [24]. For the regression task being examined here the variance of the ensemble predictions can be used. The detection mechanism is

$$\frac{1}{M}\sum_{i=1}^{M}\left[\mathcal{F}(\hat{\boldsymbol{w}}_{1:n+k}; \boldsymbol{\theta}^{(i)})\right]^2 - \left[\frac{1}{M}\sum_{i=1}^{M}\mathcal{F}(\hat{\boldsymbol{w}}_{1:n+k}; \boldsymbol{\theta}^{(i)})\right]^2 > \beta$$

In this work, where deep-learning approaches are used, a simple ensemble can be generated by using different seeds to randomly initialise the model training and a simple average to obtain an ensemble prediction.

For SLA one of the standard approaches to detecting malpractice, as well as detecting when a candidate cannot generate an appropriate response to a prompt, is to use off-topic response detection [26]. If a candidate appends an optimal adversarial attack phrase to the end of a valid response to the prompt, this may impact the relevance of the response to the prompt. Thus the detection mechanism is based on

$$P(\texttt{rel}|\boldsymbol{p}, \hat{\boldsymbol{w}}_{1:n+k}) < \beta \tag{5}$$

where $\boldsymbol{p}$ is the prompt associated with the response. Here a hierarchical attention based topic model (HATM) for off-topic spontaneous response detection [33] was used.

As previously discussed, two forms of SLA systems can be considered, a neural-based approach and systems based on expert features. Feature-based approaches are expected to be less sensitive to adversarial attacks. Thus differences in performance between the two systems can be used for attack detection. Considering a feature-based system using a GP-based grader, attack detection is based on:

$$\mathcal{F}(\hat{\boldsymbol{w}}_{1:n+k}; \boldsymbol{\theta}) - \texttt{map}(\mathcal{F}_{\texttt{gp}}(\boldsymbol{\phi}(\hat{\boldsymbol{w}}_{1:n+k}))) > \beta \tag{6}$$

where $\boldsymbol{\phi}()$ is the feature extraction process for the word-sequence. Rather than using the raw predicted GP-score, a mapped version is used based on a linear mapping, $\texttt{map}()$ from the GP-score to the neural assessment score estimated on a held-out data set. This should handle, for example, the mismatch in the average scores from the two forms of grader in Table 2. For each detection scheme, the threshold β is varied to generate precision-recall curves.

3. Experiments

Experiments were run on answers to the Linguaskill-Business (L-Bus) Use of Business English test [34], where the candidate responds to prompts from five different groups (sections A-E)

[1] Exactly how the adversarial output from the ASR system is produced is not considered in this paper. It is possible that the candidate could simply speak the word sequence, assuming that the ASR is accurate, or the ASR system itself may be adversarially attacked.

[2] Here knowledge of the vocabulary is assumed. In practice provided the selected subset is large enough this knowledge is not necessary.

[3] The grade-dependent LM was implemented using the CUED-RNNLM v1.1 toolkit [32]

to predict the final grade. The training and test data consists of non-native English spoken by candidates from 6 L1s (first language). A held-out evaluation set of 202 speakers, approximately balanced for L1 and across the CEFR grades (A1-C[4]) was used for testing. For this data the ASR system had an average word error rate of 19.5%. Reference scores were provided by expert graders. The graders were trained on a set of ∼900 speakers from the same set of L1s, using operational grader reference scores. A held-out subset of 200 speakers, balanced for grade and L1, was used to determine the adversarial phrases.

In this work all graders were constructed to predict scores for each of the five sections, then the scores averaged to yield the final score. Two feature-based graders were built; one GP-based [35] (GP_{txt}) and the other DNN-based [36] (DNN_{txt}). The features for these systems were the text features described in [35]. For the neural assessment system (Neur_{txt})[5], BERT was used to extract the word-embeddings, followed by a multi-head-self attention mechanism [37]. The output of this process was then fed to the same DNN configuration as [36]. For the neural systems an ensemble of 10 models were built and the predictions averaged to yield the final score.

To confirm that the most important features were text-based, motivating attacking these in the SLA, three deep neural assessment systems were examined: the text based system (Neur_{txt}); a deep pronunciation system [38] ($\text{Neur}_{\text{pron}}$) and a rhythm based system [39] ($\text{Neur}_{\text{rytm}}$). Table 1 presents the performance of these neural-based systems, as well as a text only (GP_{txt}) and all [35] (GP_{all}) feature-based GP system on the overall test performance (the average over all sections A-E). The system is assessed in terms of Pearson Correlation Coefficient (PCC) and RMSE to the expert scores, as well as the percentage of predictions within half (<0.5) and one (<1.0) grade level of the expert score.

Table 1: *Baseline Performance of, sections A-E, graders.*

System	PCC	RMSE	<0.5	<1.0
Neur_{txt}	0.878	0.587	66.8	91.4
$\text{Neur}_{\text{pron}}$	0.819	0.699	54.1	85.5
$\text{Neur}_{\text{rytm}}$	0.815	0.697	55.9	86.4
$\text{Neur}_{\{\text{txt}\oplus\text{pron}\oplus\text{rytm}\}}$ $\alpha = [0.83, 0.04, 0.13]$	0.884	0.581	67.0	90.5
GP_{txt}	0.855	0.643	60.4	87.7
GP_{all}	0.881	0.606	60.5	91.4

Table 1 clearly shows that for both the feature and ensemble neural-based systems, the text features are the most important. The optimal linear combination of the neural systems gives a weight of 0.83 to the prediction from the text system. The table also illustrates that pure neural systems are highly competitive with expert derived systems, without the need to define features. Comparing the GP and neural systems with text only features shows that more information can be extracted by the neural approach than hand derived features. Given the dominance of the text features in performance, text-based adversarial attacks are of most relevance to SLA.

As each of the sections of the test have different attributes, and different neural assessment systems, it is sensible to generate a different universal attack for each section. This work focused on section C where candidates can talk for up to 60 seconds on a prompted topic. The average response length was 41 seconds of speech. Table 2 shows the baseline performance of the text deep neural model and the feature-based GP model. In addition to the ensemble performance the single system performance is also given for the neural system. To construct the adversarial attack only one of the members of the ensemble was used. It was found that this attack transferred to all members of the ensemble, as expected from the relatively small standard deviation. The table also shows that the neural and GP systems are complementary.

Table 2: *Baseline performance (on section C) of the text-based GP and Neural graders. ± indicates the standard deviation.*

Grader	Score	PCC	RMSE	<0.5	<1.0
GP_{txt}	3.88	0.749	0.786	54.0	80.7
Neur_{txt}	3.49 ± 0.14	0.744 ± 0.01	0.818 ± 0.06	48.9 ± 6.55	79.4 ± 2.86
-ensemble	3.49	0.749	0.727	59.9	83.2
$\text{GP}_{\text{txt}}\oplus\text{Neur}_{\text{txt}}$	3.69	0.774	0.678	61.4	83.7

The substitute model [8], used to obtain the subset vocabulary needed for the discrete greedy search adversarial attack of the text-based graders, employed a different text-based architecture for grading. The embedding stage used a simple, context independent word2vec [40] transformation. The 100 most effective adversarial attack words for the substitute system were then used for the subset vocabulary, as described in section 2.

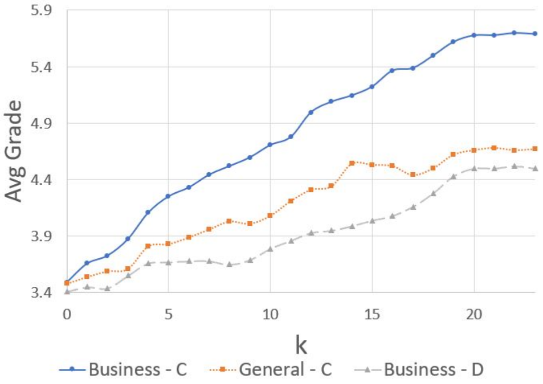

Figure 1: *Transferability of k-word attack phrase found for the Neural model trained on L-Bus, section C*

The text-based neural grader for section C of the Linguaskill business test was then adversarially attacked using the approach in section 2. The change in the average score as the number of words in the universal phrase is increased is shown in figure 1. The average response length for this section was 79 words. After adding a 20 word phrase the performance[6] has almost saturated at 5.7 out of 6 in terms of the average system score over all test speakers. It is also of interest to see how this universal attack performs on a task with different types of topics in the prompts (the Linguaskill general test with average response length 83 words) and a different form of prompt question (section D average response length 85 words). Though for both systems the average score is increased, the increase is approximately half that of the matched attack.

[4] Due to limited data grades C1 and C2 are combined.

[5] Available at: `https://github.com/rainavyas/NeurTxtGrader`

[6] As a sanity check random k length "phrases" from the vocabulary were also used. As expected these did not improve the average score.

Table 3: *Impact of the 6 word Neural adversarial attack* NEUR*-adv or GP adversarial attack* GP*-adv on different graders*

Grader (+adv)	Score	PCC	RMSE	<0.5	<1.0
GP_{txt}	3.88	0.749	0.786	54.0	80.7
+ GP-adv	4.27	0.744	1.037	30.7	68.3
+ NEUR-adv	4.02	0.749	0.863	49.0	76.7
DNN_{txt}	3.70	0.750	0.732	56.9	83.7
+ GP-adv	4.23	0.691	1.038	31.7	72.3
+ NEUR-adv	3.84	0.750	0.772	53.0	82.7
Neur_{txt}	3.49	0.749	0.727	59.9	83.2
+ GP-adv	3.54	0.753	0.702	58.4	83.2
+ NEUR-adv	4.33	0.700	1.110	27.2	62.9

Rather than operating near the saturation point for the attack, a shorter attack of length 6 was used for a detailed analysis of the system. This shorter attack still yielded an increase in the average score of 0.84, and should be more challenging to detect. Table 3 shows the impact of adversarial attack phrase (NEUR-adv), on the neural assessment system and the GP text system. Though again both scores are increased, the GP-based grader only increased by 0.14, significantly less than the neural-based system. In addition Table 3 shows the impact of an adversarial attack on the GP feature-based system (GP-adv). The feature-based system is less sensitive than the neural system to adversarial attacks, with minimal transfer of the GP-based attack to the neural system. Finally an alternative DNN feature-based systems (DNN_{txt}) was also examined. This system is far more impacted by the GP-optimised attack rather than the Neural. This is expected as the same set of features are used.

From Table 3 the impact of the adversarial attack NEUR-adv on the Neural assessment system Neur_{txt} is very large, decreasing the number of candidates within half a grade point from almost 60% to less than half that, 27.2%. This motivates the use of the adversarial attack detection approaches described in section 2. As the penalty for incorrectly detecting an adversarial attack is high, the candidate may be rejected or given a score of 0, precision is more important than recall for this task. Thus $F_{0.5}$ is used to give a single point summary of the system performance.

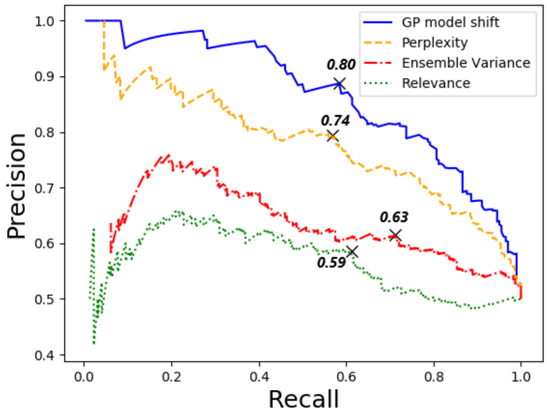

Figure 2: *Precision-Recall curves for different detection approaches for the Neural assessment system with 6* NEUR*-adv words*

Figure 2 shows the precision and recall curves and the optimal $F_{0.5}$ value for the four detection schemes from section 2 as β varies. The performance of the off-topic response detection is the worst, as only a short adversarial phrase is appended to a valid, on-topic, response. Although ensemble diversity and perplexity show reasonable performance, the best performing detection scheme is GP shift, with an $F_{0.5}$ score of 0.8.

The current adversarial attack is based on appending the adversarial attack to the end of the response. In order to assess the transferability of this attack to different positions the same 6 word phrase was appended to the beginning or to the middle of the original response for the NEUR-adv attack. This yielded average score values of 4.08 and 4.13 respectively, compared to appending to the end of 4.33. Thus a large grade increase of greater than half a grade was possible even for these sub-optimal attacks for the neural grader.

Table 4: *Detection evasion attacks on the Neural grader*

Grader (+adv)	Score	PCC	RMSE	<0.5	<1.0
Neur_{txt}	3.49	0.749	0.727	59.9	83.2
+ NEUR-adv	4.33	0.700	1.110	27.2	62.9
+ GP-Det-adv	4.15	0.715	0.975	35.1	69.8
+ PERP-Det-adv	4.22	0.711	1.020	32.7	66.3

It is possible to attack a system with knowledge of the defence mechanism. Attacks were generated independently to evade the GP Shift (GP−Det-adv) and perplexity (PERP−Det-adv) detection processes by ensuring that the final GP shift and perplexity were less than the corresponding thresholds used to generate the $F_{0.5}$ scores in Figure 2. These attacks (Table 4) yield lower increases in the average score (4.15 and 4.22 correspondingly) than the unconstrained NEUR-adv attack. It is of course possible to operate at lower thresholds for the detection evasion attacks, or combine the detection approaches, to further reduce the impact of adversarial attacks.

4. Conclusions

This paper has examined a simple, universal black-box adversarial attack for deep-learning based spoken assessment systems. The aim is to generate a single, universal phrase that when uttered at the end of a valid response to a prompt will improve the performance of any candidate. The system is evaluated on a free-speaking section of an English assessment test, Linguaskill Business. The paper shows that spoken language assessment systems are susceptible to these universal attacks. Even a short six word phrase can yield nearly a one point increase in the average grade for the test speakers. The impact of adversarial attacks for these neural systems is compared to more traditional feature-based systems which are found to be less sensitive to adversarial attacks. Four defence mechanisms, including the standard perplexity score, as well as assessment specific schemes, are also discussed. These can accurately detect attacks, but can also be used as part of the adversarial attack generation, if the form of detection is known.

The work in this paper has focused on the system using only the text from the ASR system as this yields the most important features for spoken language assessment, and it is easy for a candidate to learn a single phrase to utter in addition to their standard response. Universal attacks on other features, such as pronunciation features, are also possible and will be examined in future work.

5. References

[1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Sig. Proc. Magazine*, vol. 29, pp. 82–97, November 2012.

[2] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," in *Proc. of Speech Synthesis Workshop (SSW)*, 2016.

[3] I. Sutskever *et al.*, "Sequence to sequence learning with neural networks," in *Proc. of 27th Int. Conf. on Neural Information Processing Systems*, 2014, pp. 3104–3112.

[4] H. Xu *et al.*, "Text classification with topic-based word embedding and convolutional neural networks," in *Proc. of 7th ACM Int. Conf. on Bioinformatics, Computational Biology, and Health Informatics*, 2016, pp. 88–97.

[5] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," *CoRR*, vol. abs/1608.04644, 2016. [Online]. Available: http://arxiv.org/abs/1608.04644

[6] A. Ilyas *et al.*, "Prior convictions: Black-box adversarial attacks with bandits and priors," *CoRR*, 2018. [Online]. Available: https://arxiv.org/pdf/1807.07978

[7] Z. Yan *et al.*, "Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks," *CoRR*, vol. abs/1906.04392, 2019. [Online]. Available: http://arxiv.org/abs/1906.04392

[8] N. Papernot *et al.*, "Practical black-box attacks against deep learning systems using adversarial examples," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1602.02697

[9] S. Moosavi-Dezfooli *et al.*, "Universal adversarial perturbations," *CoRR*, vol. abs/1610.08401, 2016. [Online]. Available: http://arxiv.org/abs/1610.08401

[10] A. Kurakin *et al.*, "Adversarial machine learning at scale," in *Proc. of International Conference on Learning Representations (ICLR)*, 2017.

[11] Y. Liu *et al.*, "Delving into transferable adversarial examples and black-box attacks," *CoRR*, vol. abs/1611.02770, 2016. [Online]. Available: http://arxiv.org/abs/1611.02770

[12] C. Xie *et al.*, "Improving transferability of adversarial examples with input diversity," *CoRR*, vol. abs/1803.06978, 2018. [Online]. Available: http://arxiv.org/abs/1803.06978

[13] N. Inkawhich *et al.*, "Feature space perturbations yield more transferable adversarial examples," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7059–7067.

[14] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," *CoRR*, vol. abs/1706.06083, 2017. [Online]. Available: http://arxiv.org/abs/1706.06083

[15] A. Kurakin *et al.*, "Adversarial examples in the physical world," in *Proc. of International Conference on Learning Representations (ICLR)*, 2017.

[16] N. Papernot *et al.*, "The limitations of deep learning in adversarial settings," in *Proc. of IEEE European Symposium on Security and Privacy (EuroS&P)*, 2015, pp. 372–387.

[17] Q. Lei *et al.*, "Discrete adversarial attacks and submodular optimization with applications to text classification," in *Proc. of Machine Learning and Systems*, 2019, pp. 146–165.

[18] W. Wang *et al.*, "Towards a robust deep neural network in texts: A survey," *CoRR*, vol. abs/1902.07285, 2019. [Online]. Available: http://arxiv.org/abs/1902.07285

[19] M. Alzantot *et al.*, "Generating natural language adversarial examples," in *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 2890–2896.

[20] J. Ebrahimi *et al.*, "Hotflip: White-box adversarial examples for NLP," *CoRR*, vol. abs/1712.06751, 2017. [Online]. Available: http://arxiv.org/abs/1712.06751

[21] P. Yang *et al.*, "Greedy attack and gumbel attack: Generating adversarial examples for discrete data," *CoRR*, vol. abs/1805.12316, 2018. [Online]. Available: http://arxiv.org/abs/1805.12316

[22] M. Behjati *et al.*, "Universal adversarial attacks on text classifiers," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7345–7349.

[23] T. Pang *et al.*, "Robust deep learning via reverse cross-entropy training and thresholding test," *CoRR*, vol. abs/1706.00633, 2017. [Online]. Available: http://arxiv.org/abs/1706.00633

[24] A. Malinin and M. J. Gales, "Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness," in *Proc. of Advances in Neural Information Processing Systems 32 (NeuRIPS)*, 2019, pp. 14 520–14 531.

[25] T. Strauss *et al.*, "Ensemble methods as a defense to adversarial perturbations against deep neural networks," 2017. [Online]. Available: http://arxiv.org/abs/1709.03423

[26] V. Raina *et al.*, "Complementary systems for off-topic spoken response detection," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Jul. 2020, pp. 41–51.

[27] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[28] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[29] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, no. 2, pp. 282–306, 2011.

[30] R. van Dalen *et al.*, "Automatically grading learners' English using a gaussian process," in *Proc. of ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, 2015.

[31] J. Devlin *et al.*, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.

[32] X. Chen *et al.*, "CUED-RNNLM – an open-source toolkit for efficient training and evaluation of recurrent neural network language models," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[33] A. Malinin *et al.*, "A hierarchical attention based model for off-topic spontaneous spoken response detection," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 397–403.

[34] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

[35] Y. Wang *et al.*, "Towards automatic assessment of spontaneous spoken english," *Speech Communication*, vol. 104, pp. 47–56, 2018.

[36] A. Malinin *et al.*, "Incorporating uncertainty into deep learning for spoken language assessment," in *Proc. of 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 45–50. [Online]. Available: https://doi.org/10.18653/v1/P17-2008

[37] A. Vaswani *et al.*, "Attention is all you need," in *Proc. of 31st International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.

[38] K. Kyriakopoulos *et al.*, "A deep learning approach to assessing non-native pronunciation of english using phone distances," in *Proc. of INTERSPEECH*, 2018, pp. 1626–1630.

[39] K. Kyriakopoulos, K. M. Knill, and M. J. F. Gales, "A deep learning approach to automatic characterisation of rhythm in non-native English speech," in *Proc. of INTERSPEECH*, 2019, pp. 1836–1840.

[40] T. Mikolov *et al.*, "Efficient Estimation of Word Representations in Vector Space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781