ROBUSTNESS TO ADVERSARIAL ATTACKS AND SCALING LAW

Rodwell Nicolas Bent
rodwell-nicolas.bent@polymtl.ca

Tom Marty
tom.marty@umontreal.ca

Rafael Hernandez Garcia
rafael.hernandez.garcia@umontreal.ca

Siddhika Arunachalam
siddhika.arunachalam@umontreal.ca

ABSTRACT

In this paper, we aim to show, regarding deep learning models, that there exists a relationship between the loss and the number of parameters. Ultimately, uncovering the emergence of power laws and testing the robustness of language model with scale. We evaluate the performance of various GPT models such as GPT-Neo 125M, GPT-Neo 1.3B, GPT-Neo 2.7B, GPT-J 6B against adversarial attacks. Current literature on adversarial attacks do not generally quantify how vulnerable these models are to attacks. The two adversarial attacks namely BAE (BERT-based Adversarial Examples) and TextFooler are conisidered as a means for Text Classification. Therefore, this piqued our interest and suggested we should dive deeper into the subject. We will train these GPT models on the RottenTomatoes dataset to evaluate the effects of scale on adversarial robustness.

1 Introduction

Recently, it has been established that the loss of neural networks follow power laws which scale with the size of the data or network [7]. This has been demonstrated for natural language processing models [9], including neural translation [6]. The exponents of the power law controls the speed of the improvement with regard to different parameters. Neural language laws also shows that large models are sample efficient and that with a fixed compute budget it is better to train a large model and stop before it has fully converged [9]. Improvement in loss is not the only way to observe model performance, and there has been a significant body of work on the robustness of neural models to adversarial attacks [17]. There have been various ways to quantify robustness [2, 19, 10]. It has also been shown that different networks trained on the same data set have fallen victim to the same type of adversarial examples [15]. This leads to the question does robustness follow scaling laws.

We aim to examine this by using trained large scale language models present in the hugging face library trained by Eleuther AI [1]. The GPT-Neo and GPT-J models offer a way to measure the robustness of models that have different numbers of parameters, from 125M-6B. Most of the measures of robustness correspond to measures of accuracy, so they are not a good way to judge neural scaling laws, instead we will build an adversarial dataset. Once the dataset has been created, we can measure the overall loss of each model and verify if the robustness of language models follow scaling laws. This could also be looked at as a special case of the effect of model size on worst group generalization, where the worst group is an adversarial group [14].

2 Motivation

Massive natural language models have become common place in the modern machine learning landscape. At the same time, the robustness of models to adverserial examples has also become an important area of research. It is therefore important to examine how scale impacts robustness and whether all of these larger models are able to improve their robustness.

3 Related Work

3.1 Adversarial Robustness

A Voronoi-epsilon adversary is proposed in [10] which the two notions of perturbation helps to manage the adversary. Because of this, a trade-off is not produced between accuracy and adversarial accuracy, even when ϵ is large due to adversarial accuracy based on this adversary.

A gradient-guided search over tokens that finds trigger sequences which are short is proposed in [15]. Also, the triggers that are optimized for specific models can be used for the other models and all the tasks considered.

An adversarial training process named as Attacking to Training (A2T) is introduced in [**?**]. Datasets such as Rotten Tomatoes, IMDB, Yelp and SNLI are used for training purposes which uses models like BERT and RoBERTa. The authors also proved that the accuracy, cross-domain generalization and interpretability can be improved in NLP models.

3.2 Scale Laws

There has been a lot of research on creating adversarial examples and hardening adversarial, but it lacks research in the area of performance measures for evaluating adversarial robustness. By taking inspiration from this, the study in [2] introduces residual error for evaluating the adversarial robustness at each sample of a deep neural network and also works to differentiate between the examples of adversarial and non-adversarial. The authors concluded that the proposed concept was critical in the design of robust models that are adversarial.

In terms of cross-entropy loss, a certain scaling law is observed as a function of model size in [6]. A formula is presented which describes the scaling behavior of cross-entropy loss. The differences between the encoder and decoder scaling shows different power law exponents when observed. Also, the connection between the cross-entropy loss and quality of the translations that are generated are studied.

The training sets growth leads to the growth of generalization error and model size. The methodology is experimented on four different Machine Learning experiments such as machine translation, image processing, speech Recognition and language modeling. The authors in paper [7] confirms that the accuracy of the Deep Learning model as a power law improves as it grows. In addition, changes to the model architecture and optimizer within each domain only shift the learning curve. However, it does not affect the power law index.

Empirical scaling laws for language model performance on the cross-entropy loss is investigated in [9]. Loss is scaled as a power law depending on model size, dataset size, and compute power used for training. Also, an interesting point which the authors figured is that the dependence of overfitting on the dataset size and the dependence of training speed on the model size can be controlled with the help of simple equations. Another important finding is that the sample efficiency in larger models are significantly more, such that the compute-efficient training comprises training very large models on relatively small amount of data.

The effect of model size on worst-group generalization under empirical risk minimization (ERM) is tested in various settings with respect to architectures, domain and model sizes. It was also studied in [14] that by increasing the model size of pre-trained models consistently, the performance on datasets like MultiNLI and Waterbirds also improved.

4 Methodology and Datasets

Our aim is to study the impact of adding adversarially generated samples to the training for models of different sizes. Specifically, we expect to find a log-linear empirical scaling law [9] between the size of the language models and the loss on the evaluation set for text classification task. We used the TextAttack library [12] to automate the training process with or without adversarial data. This library allows to use very easily models and datasets from big machine learning libraries such as Hugging face and also contains many adversarial attacks called recipe already implemented.

We have chosen to use the following language models gently provided by Eleuther AI : GPT-Neo 125M [3], GPT-Neo 1.3B [3], GPT-Neo 2.7B [3], GPT-J 6B [16], which have the particularity of having been pre-trained on the same dataset The Pile [4](825 gb), which is a necessary property to guarantee the relevance of the comparison.

The size of the models ranges from 125 million parameters to 6 billion parameters, allowing us to study the impact of the model scale on more than one order of magnitude. Our largest models are comparable in size to current state-of-the-art models.

These large models are capable of encoding high-level concepts and have a fairly good idea of the overall syntactic and semantic coherence of a text. Therefore, we decided to limit our study to word-level adversarial attacks, which act by changing/adding one or more words to the sequence. These methods generally produce sentences that are plausible and semantically close to the original sentence. The generated sentences are more difficult to detect by conventional language models, which is what we are looking for in order to efficiently benchmark our models.

We have limited ourselves to 2 different close to the state-of-the-art adversarial attacks:

- **BAE**, BERT-based Adversarial Examples for Text Classification [5]: which is a conventional BERT language model, it replaces and inserts tokens in the original text by masking a portion of the text and leveraging the BERT model to generate alternatives for the masked tokens.

- **TextFooler** [8]: it assigns an importance score to the different words in a sentence, and replaces the most important words from a vocabulary generated in such a way as to maintain semantic similarity.

As mentioned earlier, we were interested in the task of text sentiment classification. To do so, we trained and evaluated our models on the **RottenTomatoes** [13] dataset available on Hugging Face, which contains 5,331 positive and 5,331 negative processed sentences from Rotten Tomatoes movie reviews. We could also have AmazonPolarities and IMDB [11] datasets to validate our results obtained on the first dataset. We decided to focus on a single dataset to be able to train more models given our limited computing power.

text (string)	label (class label)
the rock is destined to be the 21st century's new " conan " and that he's going to make a splash even greater than arnold schwarzenegger , jean-…	1 (pos)
the gorgeously elaborate continuation of " the lord of the rings " trilogy is so huge that a column of words cannot adequately describe co-…	1 (pos)
effective but too-tepid biopic	1 (pos)
if you sometimes like to go to the movies to have fun , wasabi is a good place to start .	1 (pos)
emerges as something rare , an issue movie that's so honest and keenly observed that it doesn't feel like one .	1 (pos)

Figure 1: Samples from the **RottenTomatoes** dataset and their associated label/sentiment

The different models are trained on the augmented-RottenTomatoes dataset with increasing added adversarial samples (No samples / 10 / 100 / 1000 / 5000) on 3 full epochs with a learning rate of 5e-5.

After every epochs, all models are evaluated on a fixed dataset containing 1250 adversarial samples, 625 generated using each attack method (Textfooler and BAE). The metric being used for the

empirical scaling law is the evaluation loss on this dataset for each models. The entire pipeline is recapitulated in Fig. 2.

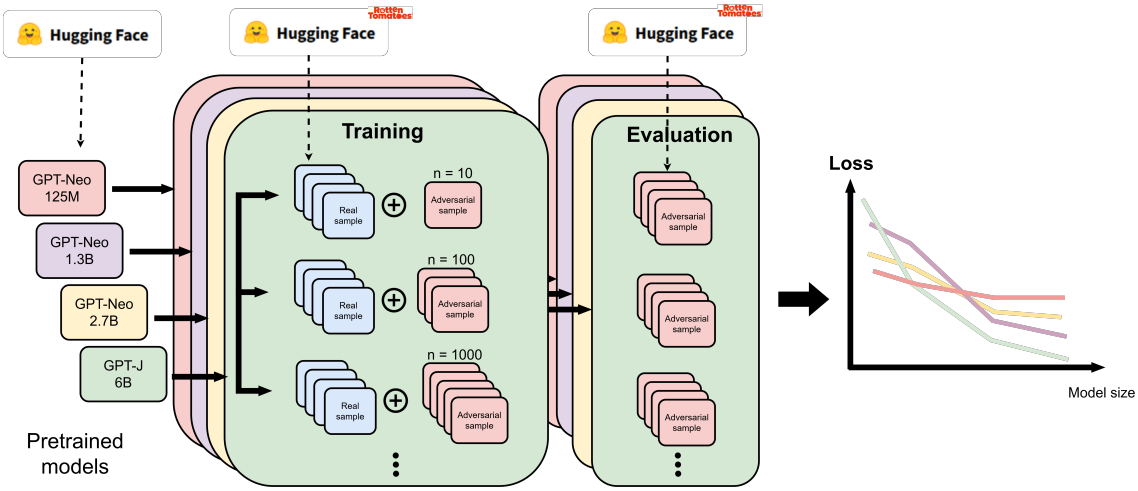

Figure 2: Summary of the whole learning and evaluation pipeline.

Although the loss is the metric commonly used for model comparison, in a context of sentiment classification we also consider the attack success rate (proportion of samples classified by the agent out of all adversarial samples) to ensure that a decreasing loss is associated an increasing classification performance.

Finally, we were interested in the L2-distance between the sentence embeddings computed by the transform output models just before the dense classification layer. We expect to observe a correlation between the performance of a model and its ability to keep the embeddings of a sample and its version modified by the adversarial attack relatively close to each other.

5 Results and Discussion

To determine whether scale impacts adversarial robustness we examined different measures of robustness for different models at scale. First, using our adversarial dataset, we measured the average adversarial loss for all the different sizes of models Figure 3. The results show that in this narrow range the adversarial loss behaves like a scale law. The scale law follows $L = 2.134N^{(-.0579)}$ where N is the number of parameters.

Subsequently we analyzed the attack success rate on the trained model Figure 4 shows the average success rate and Figure 5 shows the success rate on models trained on the individual models. The attack success rate decreases as a function of model size but only very slightly. Meanwhile, for the individual models, attack success rate is not a monotonically decreasing function of amount of adversarial data the model was trained on. For all models, 100 adversarial data examples gave the lowest loss. These results align with Figure 2 in [18]

Finally, we examine the L2 distance between the original text and the adversarial text in Figure 6. The decreasing relationship between model size and L2 distance show that the models are not as impacted by the adversarial examples. These results also align with Figure 3 [18], which demonstrate that more robust models have smaller L2 distances.

All of these results are preliminary and more work has to be done to verify, we will try to outline some of the strengths and weakness of the results. First, despite the fact that a scale law was calculated for adversarial loss it was done in a very narrow band of model size. Many more models would have to be trained at different sizes, larger and smaller to see if there are any phase transitions. Second, only one size of data was used so it would be informative to examine how dataset size impacts adversarial robustness as well. Additionally, looking at FLOPS and not just model size can give more information on how scale impacts robustness. Furthermore, we only trained models on one dataset and it would be imperative to extend these results to different datasets. Finally, we have only used an adversarial

dataset as a measure of adversarial robustness but there are other measures of adversarial robustness and exploring those would be beneficial.

In spite of these weaknesses, these results are still quite promising and do demonstrate that there is a strong relationship with model size and adversarial robustness. Another strength of this paper is that it's results corroborate some results in [18].

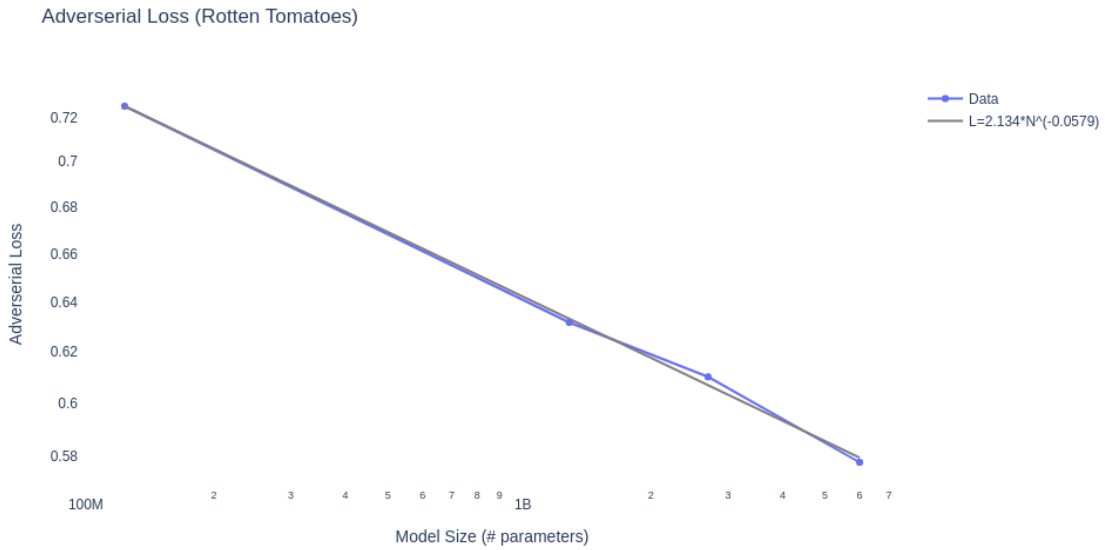

Figure 3: Average loss on adversarial examples for different model sizes in log log scale. Each dot represents an average of the losses for all models of the same size. Different models of the same size were trained with different amounts of adversarial data introduced during training.

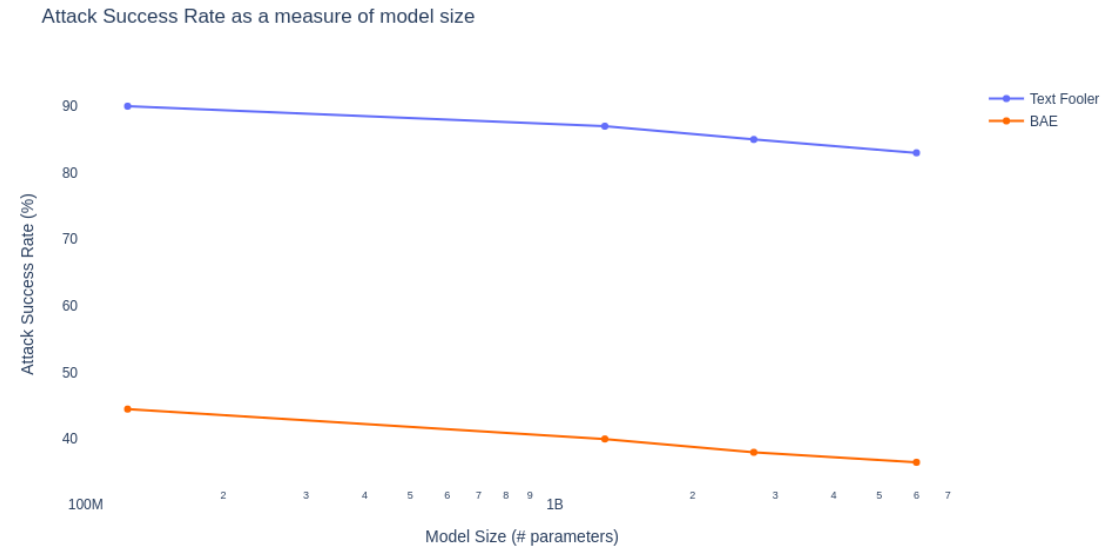

Figure 4: Average attack success rate of two different attacks, TextFooler [] and BAE []. The graph is in log scale on the x axis.

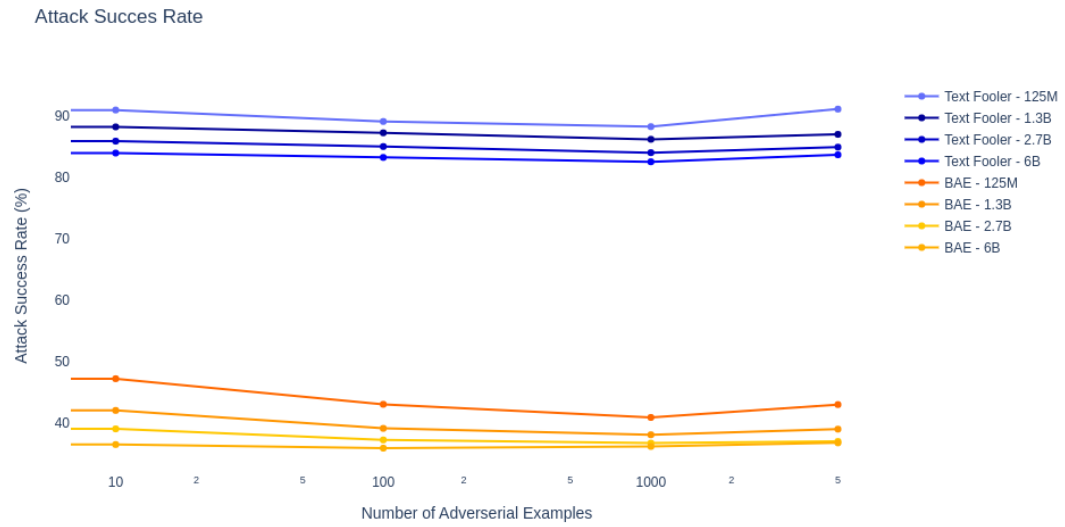

Figure 5: Attack success rate on different attacks, TextFooler [] and BAE [] for models trained on different amounts of adversarial data.

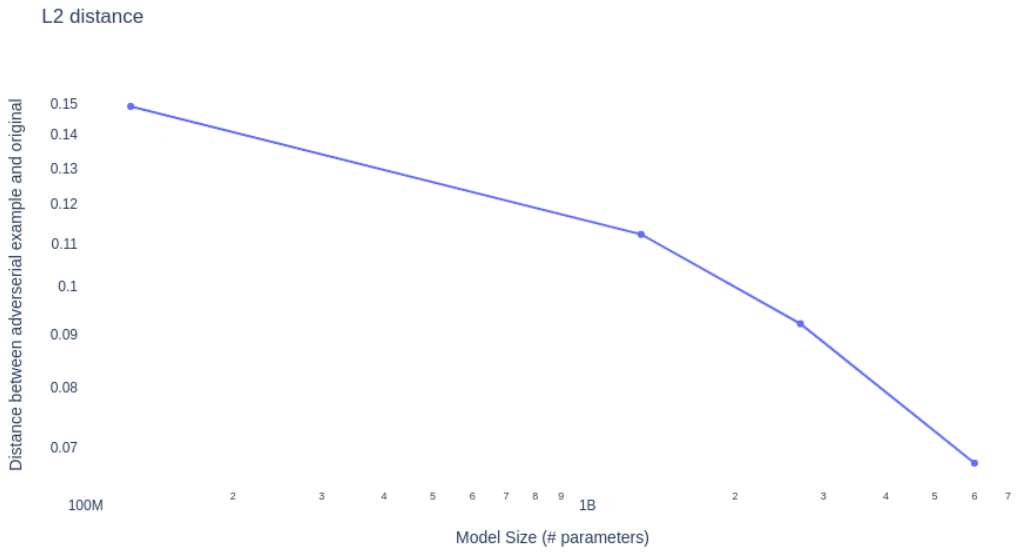

Figure 6: L2 distance between the representation of the original text and the adversarial text. This shows a monotonically decreasing relationship between L2 distance and model size.

6 Conclusion and Future Work

Our current results, albeit still preliminary, suggests indeed that scale improves adversarial robustness. For the different sized of GPT models used like GPT-Neo 125M, GPT-Neo 1.3B, GPT-Neo 2.7B, GPT-J 6B, the average adversarial loss obtained shows behavior like that of a scale law. The attack success rate shows that the two word-level adversarial attacks follow a similar pattern and decrease marginally as the model size (Number of parameters) increase. Furthermore, the L2 distance between the adversarial and original example declines as the model size increases indicating that the L2 distances are less when the number of robust models increases. All of these give evidence that scale

is important for adversarial robustness but more work has to be be done to solidify the evidence. For future work it would be important to expand the experiments we have done. First, we should expand the scale of our models, we have only looked at a narrow band of 125M-6B parameters and we should look at both larger like the 20B and 175B models as well as much smaller models. Second, we should look at training initial training of smaller datasets and see how dataset size affects scale. Third, we should look at robustness as a measure of FLOPs. Subsequently, we should try and find different adversarial robustness measures. Finally, we need to do our final training on multiple different datasests like IMBD or Amazon Polarities. Once all of these experiments have been done we will have a much clearer picture on how scale impacts adversarial robustness.

References

[1] Eluether ai large scale gpt models. `https://huggingface.co/EleutherAI`. Accessed: 2022-03-04.

[2] ABOUTALEBI, H., SHAFIEE, M. J., KARG, M., SCHARFENBERGER, C., AND WONG, A. Residual error: a new performance measure for adversarial robustness. *arXiv preprint arXiv:2106.10212* (2021).

[3] BLACK, S., LEO, G., WANG, P., LEAHY, C., AND BIDERMAN, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, Mar. 2021. If you use this software, please cite it using these metadata.

[4] GAO, L., BIDERMAN, S., BLACK, S., GOLDING, L., HOPPE, T., FOSTER, C., PHANG, J., HE, H., THITE, A., NABESHIMA, N., PRESSER, S., AND LEAHY, C. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[5] GARG, S., AND RAMAKRISHNAN, G. Bae: Bert-based adversarial examples for text classification, 2020.

[6] GHORBANI, B., FIRAT, O., FREITAG, M., BAPNA, A., KRIKUN, M., GARCIA, X., CHELBA, C., AND CHERRY, C. Scaling laws for neural machine translation. *arXiv preprint arXiv:2109.07740* (2021).

[7] HESTNESS, J., NARANG, S., ARDALANI, N., DIAMOS, G., JUN, H., KIANINEJAD, H., PATWARY, M., ALI, M., YANG, Y., AND ZHOU, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409* (2017).

[8] JIN, D., JIN, Z., ZHOU, J. T., AND SZOLOVITS, P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment, 2019.

[9] KAPLAN, J., MCCANDLISH, S., HENIGHAN, T., BROWN, T. B., CHESS, B., CHILD, R., GRAY, S., RADFORD, A., WU, J., AND AMODEI, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[10] KIM, H., PARVIAINEN, P., AND MALDE, K. Measuring adversarial robustness using a voronoi-epsilon adversary. *arXiv preprint arXiv:2005.02540* (2020).

[11] MAAS, A. L., DALY, R. E., PHAM, P. T., HUANG, D., NG, A. Y., AND POTTS, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 2011), Association for Computational Linguistics, pp. 142–150.

[12] MORRIS, J. X., LIFLAND, E., YOO, J. Y., GRIGSBY, J., JIN, D., AND QI, Y. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp, 2020.

[13] PANG, B., AND LEE, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL* (2005).

[14] PHAM, A., CHAN, E., SRIVATSA, V., GHOSH, D., YANG, Y., YU, Y., ZHONG, R., GONZALEZ, J. E., AND STEINHARDT, J. The effect of model size on worst-group generalization. *arXiv preprint arXiv:2112.04094* (2021).

[15] WALLACE, E., FENG, S., KANDPAL, N., GARDNER, M., AND SINGH, S. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 2153–2162.

[16] WANG, B., AND KOMATSUZAKI, A. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[17] WANG, Y., MA, X., BAILEY, J., YI, J., ZHOU, B., AND GU, Q. On the convergence and robustness of adversarial training. In *Proceedings of the 36th International Conference on Machine Learning* (09–15 Jun 2019), K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97 of *Proceedings of Machine Learning Research*, PMLR, pp. 6586–6595.

[18] YOO, J. Y., AND QI, Y. Towards improving adversarial training of nlp models, 2021.

[19] YU, F., QIN, Z., LIU, C., ZHAO, L., WANG, Y., AND CHEN, X. Interpreting and evaluating neural network robustness. *arXiv preprint arXiv:1905.04270* (2019).