Using Adversarial Examples to Capture Psychological Representations from Deep Neural Networks

Sanket Agrawal,[1] **Abhishek Jain,**[2] **Shashi Kant Gupta** [3]

[1] Department of Mathematics & Statistics, [2] Department of Mechanical Engineering, [3] Department of Electrical Engineering
Indian Institute of Technology Kanpur, India
C-513, Hall 13, IIT Kanpur, India, Ph - (91)8800705755
sanket@iitk.ac.in, jainabhi@iitk.ac.in, shashikg@iitk.ac.in

Abstract

Deep neural networks have become immensely popular in solving computer vision task, often surpassing the human-level performance. A lot of recent research has focused on finding the similarity between DNN and human representations. Ability to extract psychological representations of an image can help us in modeling human behavior. Also, developing better insights into deep learning models can help us develop better generalization in deep learning models. Previously, researchers have attempted to extract psychological representations based on a human similarity judgment experiment. We tested one such method against adversarial images (images that look similar to original images but fools the model to miss-classify them). We found that correlation scores decreased on adversarial images resulting into a moderate correlation score. We then retrained their similarity model using adversarial images and found that this generalizes well to human similarity judgment with a strong correlation for various perturbation strength. Our work results in two interesting findings: 1. Deep learning model are indeed capable of capturing human similarity judgement and 2. Adversarial perturbations consist of human indistinguishable feature but good feature for deep learning models.

Introduction

Peterson, Abbott, and Griffiths (2016) proposed a method to adapt feature vectors learned by a DNN model to capture psychological representations using the data on human similarity judgements. In particular for the VGG16 model, they suggested considering the second last layer of the model as the learnt feature vector for any image and calculating a similarity score between any two images by taking a weighted inner product of corresponding feature vectors. They showed that when the weights were learnt by performing a Ridge regression with human data as the response variable, the performance, as measured by the square of the correlation between human similarity scores and similarity scores calculated by DNN, was much greater than when the weights were taken to be equal to one. From this result, they claimed that DNNs do learn some psychological representations, the information about which is often hidden when the features are in their raw form. However, performing a linear transformation on these feature vectors can shed some light on this information as they saw from their results.

Now, if we assume that DNNs do encode some information about psychological representations, we would not want it to be destroyed when some imperceptibly tiny amount of perturbation is added to the images or in other words, such information or in fact the features contributing to this information shall remain robust against the carefully generated adversarial examples often aimed at shattering a DNN model. This requirement is natural, since adding a perturbation that is imperceptible to humans shall not affect their psychological representations.

In this work, we show how the method suggested in (Peterson, Abbott, and Griffiths 2016) to recover information on psychological representations from DNN features is affected under the presence of adversarial attacks and how its performance decreases with an increase in the perturbation strength. We try to give a heuristic explanation for this behaviour. And based on this explanation, we attempt to suggest how these adversarial attacks can be utilized to capture more robust information about psychological representations. At the same time we also justify the results introduced by Andrew et al. that adversarial perturbations consists of human indistinguishable features but use full features for machine learning models.

Methods

We used the *Fast Gradient Sign Method* (FGSM) (Goodfellow, Shlens, and Szegedy 2014) to generate perturbations for our data. As the name suggests, FGSM is a fast method for generating adversarial images using sign of the gradient function. Formally, let θ be the parameter vector, x an input image, y be the truth label for x and $J(\theta, x, y)$ be the loss function associated with the learning task, then an optimized perturbation can be obtained as, $\eta = \epsilon \text{sign}(\nabla_x J(\theta, x, y))$ where ϵ can be varied as per the required perturbation. The corresponding adversarial example x' is then obtained as $x' = x + \eta$.

Same set of images and human similarity data as used by Peterson, Abbott, and Griffiths for their study was used for this study which is publicly available on their GitHub repository (total 120 images from an animal database). We first adapted feature vectors obtained from the original images using the method suggested in the referenced paper i.e. the similarity s_{ij} between objects i and j is modeled as in Eq. 1.

$$s_{ij} = \sum_{k=1}^{N_f} w_k f_{ik} f_{jk} \qquad (1)$$

Where f_{ik} is the k^{th} feature of image i extracted using pretrained deep learning models and w_k is its weight. The squared error in reconstructing the human similarity judgments was minimized by ridge regression. The regularization parameter for the ridge regression was obtained by performing a line search over different candidate values and was kept fixed for all other regressions that followed. Weights learned by this adaptation of original feature vectors were then stored for subsequent use. Then for each level of perturbation (as identified by ϵ), we generated adversarial examples for all 120 images and calculated the performance of the raw feature vectors obtained from them. To differentiate between these feature vectors and those obtained from natural images, we called these feature vectors as adversarial feature vectors and the former as original feature vectors. We used the square of the correlation coefficient between human similarity scores and DNN similarity scores as our performance metric. And the reported scores for each task are the the average scores over the validation set of 6-fold cross validation over the entire dataset. After raw features, we adapted these adversarial feature vectors again using the same method. Obtained weights and performance scores were noted down for each ϵ. Finally, we used the weights obtained for original feature vectors onto the adversarial feature vectors and vice versa to see how the adaptation on one performs on the other.

Results and Discussions

Figure 1 summarizes the results of our analysis. Comparing the performance of raw features (blue line) with that of the features after adaptation (orange line), we saw that the performance after adaptation on adversarial images strongly correlates with the human similarity judgement. This indicates that even though we destroy some of the features using adversarial perturbations the extracted features can very well align themselves to human judgements. The performance of adversarial feature vectors using weights learned from original feature vectors (green line) fell off at a much faster rate than all other curves, almost approaching the level of raw performance for higher magnitudes of perturbation. This indicates that adaptation weights learned using original images depends a lot on the adversarial perturbations and since these perturbations doesn't make any perceptual difference for humans, these should not be used when our ultimate aim is to extract psychological representations. Finally the red curve of the plot, representing the performance of original features using weights learned by adapting adversarial features at different levels of perturbation is of prime interest. As one can see, this curve had the slowest decreasing rate and also remained higher than all other curves, indicating a better performance at each level. In numbers, the curve started at the same point as others, i.e. at 0.77 for zero perturbation and slowly decreased to a fairly high score of 0.53 at $\epsilon = 5$. Even for $\epsilon = 20$, the performance score was about 0.44 which was comparable to the raw score for nil perturbation and the adapted score for $\epsilon = 5$. An interesting observation made from this behaviour was that the weights learnt for adversarial features performed better when they were used on original features, which indicates that when we learn adaptation weights on adversarial images the weights learns to ignore those adversarial perturbation and use only robust features of the images to adapt to human similarity judgement.

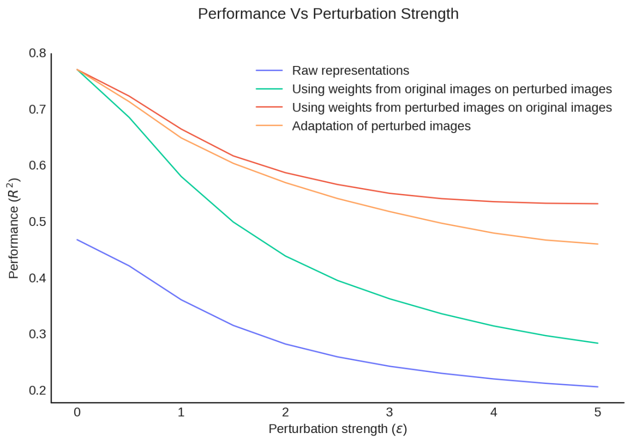

Figure 1: Performance comparison for different levels of perturbation. Raw means weights given to each feature vectors are equal. While adapted means these weights were learned using ridge regression to capture human similarity judgement.

Conclusion

Based on the results we conclude that deep learning model trained on object classification tasks are indeed capable of capturing much of the human similarity judgement if the adaptations are learned using adversarial training. At the same time, the difference in the green and red curve justify that adversarial perturbations are human indistinguishable features but are good features for machine learning model (which was previously explained by Ilyas et al. from a different approach, detailed discussion in supplementary). However we acknowledge the fact that our arguments are not based on very firm grounds, our results are transparent and reproducible and can in fact serve as a motivation for further investigation into both, the nature of adversarial examples and the quest to capture psychological representations.

References

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial Examples Are Not Bugs, They Are Features.

Peterson, J. C.; Abbott, J. T.; and Griffiths, T. L. 2016. Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164* .