Fooling Computer Vision into Inferring the Wrong Body Mass Index

Owen Levin, Zihang Meng, Vikas Singh, and Xiaojin Zhu
University of Wisconsin–Madison

ABSTRACT

Recently it's been shown that neural networks can use images of human faces to accurately predict Body Mass Index (BMI), a widely used health indicator. In this paper we demonstrate that a neural network performing BMI inference is indeed vulnerable to test-time adversarial attacks. This extends test-time adversarial attacks from classification tasks to regression. The application we highlight is BMI inference in the insurance industry, where such adversarial attacks imply a danger of insurance fraud.

1 INTRODUCTION

Body Mass Index (BMI) is a widely used health quantity calculated as kg/m^2. The world health organization categorizes BMI broadly into Underweight $[0, 18.5)$, Normal $[18.5, 25)$, Overweight $[25, 30)$, and Obese $[30, +\infty)$ [11]. Kocabey et al. recently developed a regression task Face-to-BMI [6], where they accurately predicted BMI from images of human faces. The motivation for their study was identifying how an individual's BMI affects their treatment by others on social media platforms [7].

In this paper we instead focus on the application of Face-to-BMI in the insurance industry, where adversarial attacks could become a issue. Suppose an insurance company uses a neural network to predict the BMI of their clients from photos and then uses this information to influence coverage. There are two scenarios in which an adversarial attacker may want to manipulate the input photo inperceptibly to attack the BMI predictor: (1) the attacker may want to make someone appear healthier to lower their rates; (2) conversely, make someone appear unhealthy to sabotage that person's insurance application. We demonstrate that a neural network performing Face-to-BMI is indeed vulnerable to test-time adversarial attacks. This extends test-time adversarial attacks from classification tasks (e.g. [2, 4, 9, 10]) to regression.

2 ADVERSARIAL ATTACKS ON FACE-TO-BMI PREDICTION

The victim neural network $f : \mathbb{R}^{227\times227\times3} \to \mathbb{R}$ takes as input a $227 \times 227 \times 3$ face image and outputs a BMI estimate. We use Alexnet [8] layers conv1 to fc7 plus one linear layer after fc7 to perform regression.

The threat model assumes a whitebox attacker with full knowledge of the victim weights and architecture. The attacker can edit any pixels in the photo, including those not on the human. We consider targeted attacks to force f prediction into a pre-specified target range $[L, U] \subset \mathbb{R}$.

The attack formulation find the minimum perturbation δ such that for input input X, $f(X + \delta) \in [L, U]$. Both X and $X + \delta$ must be valid images with integer pixel values in 0–255. We measure perturbation by its ℓ_p norm $\|\delta\|_p$ for some $p \in (0, \infty]$ [2, 4, 9, 10]. Thus, the ideal attack solves

$$\begin{aligned} \min_{\delta\in\mathbb{R}^{227\times227\times3}} \quad & \|\delta\|_p \\ \text{subject to} \quad & L \le f(X+\delta) \le U, \text{ and} \\ & (X+\delta) \in I := \{0,\dots,255\}^{227\times227\times3}. \end{aligned} \tag{1}$$

However, this is a difficult integer program. We heuristically solve a related problem to simply find a *small enough* δ. We reformulate the attack goal as follows: $L \le f(X+\delta) \le U \Leftrightarrow \left(f(X+\delta) - \frac{U+L}{2}\right)^2 \le \left(\frac{U-L}{2}\right)^2$. We relax the integral constraint on δ and change the objective:

$$\begin{aligned} \min_{\delta\in\mathbb{R}^{227\times227\times3}} \quad & \left(f(X+\delta) - \frac{U+L}{2}\right)^2 \\ \text{subject to} \quad & (X+\delta) \in [0,255]^{227\times227\times3}. \end{aligned} \tag{2}$$

We initialize $\delta = \mathbf{0}$ and perform early-stopping as soon as $f(X + \text{Round}(\delta)) \in [L, U]$ to encourage small norm on δ.

3 EXPERIMENTS

Datasets. We use two datasets of (photo, BMI) pairs: (1) Federal Corrections Body Mass Index (FCBMI) consists of 9045 public photos at multiple federal and state corrections facilities. (2) VisualBMI dataset with 4206 photos collected by [6] from Reddit.

Training the victim network. We train the BMI prediction network with transfer-learning. We load weights pre-trained on the ILSVRC 2012 data set for the conv1 to fc7 layers of Alexnet. Then we randomly initialize the last linear layer using Xavier [3]. Finally we fine tune the entire network's weights using our own training images. We use a random subset of 7000 images in FCBMI for fine-tuning, and keep the remaining 2045 images in FCBMI and the whole VisualBMI for testing. We pre-process the images identically to in AlexNet [8]: images are converted from RGB to BGR, re-sized to $227 \times 227 \times 3$. Finally we subtract the grand mean pixel value from each pixel in the images in the training set. This means that we provide an input in $[-255, 255]^{227\times227\times3}$ to the neural network at test time. During training we use ℓ_2 loss. We use the Adam [5] optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$. The batch size is 64 and learning rate is 0.0001.

Attack implementation. To solve (2) the attacker simulates the victim by pre-pending an extra input layer with X and 1s: The attacker freezes the weights of the entire network except δ and trains the network using projected gradient descent on the objective in (2). The architecture of this implementation is shown in Figure 1 Once training is complete, the attacker takes a final projection step and rounds δ so that $(X + \delta) \in I$.

AdvML'19: Workshop on Adversarial Learning Methods for Machine Learning and Data Mining at KDD, August 5th, 2019, Anchorage, Alaska, USA.

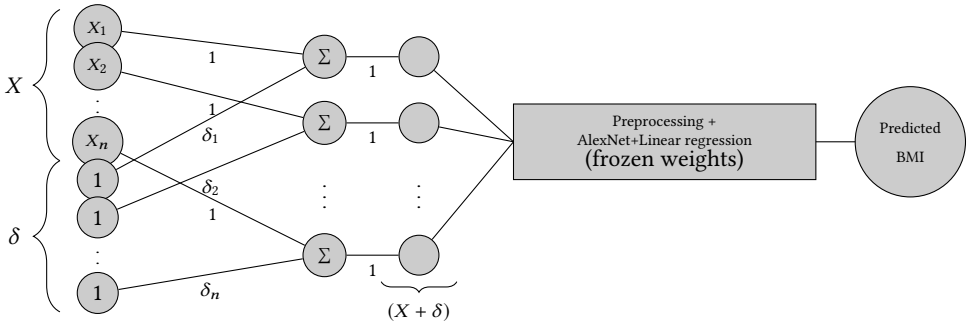

Figure 1: A cartoon of the architecture used to implement our regression attacks.

Algorithm 1 Adversarially attacking the BMI prediction network

Input: f: BMI prediction network,
X: victim image,
$K > 0$: Max iterations
Output: δ: perturbation such that $f(X + \delta) \in [L, U]$ and $(X + \delta) \in I$

$\delta \leftarrow \mathbf{0}$
$k \leftarrow 0$
while $k < K$ or $f(X + \text{Round}(\delta)) \notin [L, U]$ **do**
$\delta \leftarrow \delta - \eta_k \nabla_\delta \left(f(X + \delta) - \frac{U+L}{2} \right)^2$ {gradient descent with step size η_k}
Project δ such that $(X + \delta) \in [0, 255]^{227 \times 227 \times 3}$
$k \leftarrow k + 1$
end while
$\delta \leftarrow \text{Round}(\delta)$ {rounds δ such that $X + \delta$ is moved to the nearest point in I}
return δ {flags a failure if final δ is unsuccessful after K iterations}

Qualitative results. Figure 3 shows the BMI attack on 8 photos from the VisualBMI data set. We obscured the eyes with black boxes to preserve partial anonymity of those pictured. The boxes are not present in the original data set, so neither the prediction network nor the attacker saw or were influenced by them. Here the attack goal is to force BMI predictions into the normal range $[L, U] = [18.7, 24.9]$. The attacker succeeds at this. We note that all changes have small infinite norm: $\|\delta\|_\infty \leq 2$. Also, δs have more nonzero elements and vary more the further the original BMI is from the target range.

Quantitative results. We demonstrate two attacks separately: "make-healthy" where the attacker forces BMI predictions into $[L, U] = [18.7, 24.9]$ corresponding to normal weight, and "make-obese" with attack target range of $[L, U] = [30, 40]$ corresponding to obesity. We use the 2045 test images from the FCBMI data set and all 4206 images in the VisualBMI data set. Fig. 4 (left) shows BMI before and after attack on VisualBMI. One may expect the attack to just project the predicted BMI onto the boundary of the target range. We see almost exactly that, but there is some minor variance within the target region due to rounding of δ. Infrequently, there are large outliers where the rounding shifts the prediction to the other side of the target range. One example of this phenomenon is the right-most face in Fig. 3. Fig. 4 (right) shows $\|\delta\|_2$ under both attacks. As expected, the further a victim's initially predicted BMI from the target region, the larger the norm of the perturbation δ. Fig. 5 shows $\|\delta\|_\infty$ on the FCBMI test set. The same trend holds. Also note the maximum pixel value change is small (~5 out of 255). These attacks will be difficult for humans to perceive.

It is worth noting that the reason we see norms this high is entirely due to the rounding to integer pixels for our attacks. Without this constraint, the attack is successful with $\|\delta\|_\infty$ on the order of 10^{-2}, or roughly 100 times smaller. The success of unrounded attacks and integer attacks for various infinity norms in shown in Figure 2. In the literature, primarily FGSM-type perturbations result in integer attacks as opposed to arbitrary pixel values in the continuum of [0,255] usually seen in PGD attacks. FGSM [4] and iterative modifications of it are actually special cases of our attack when the step size η is an integer and we fix a smal maximum number of iterations. These attacks seem to result in strictly worse results than when we use a tiny step size and round.

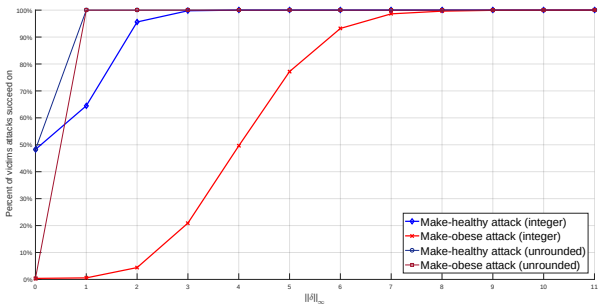

Figure 2: Portion of attacks which are successful for values of $\|\delta\|_\infty$ less than or equal to integers between 0 and 11. Note that unrounded attacks are always successful for infinity norms less than or equal to 1.

4 CONCLUSIONS AND FUTURE WORK

We have demonstrated that naïve whitebox adversarial attacks can be a threat to Face-to-BMI regression. For this reason, we urge caution when using BMI predicted from images in applications such as insurance, as they can be manipulated to make someone's rates artificially lower or higher.

The attacks in this paper requires the ability to modify any pixels. A more realistic attack would be physical, e.g. have the person wear make-up or accessories like glasses. An intermediate simulated attack could restrict the attack within face or skin pixels.

Combining these with e.g. Expectation-Over-Transformation as in [1] might allow someone to design adversarial make-up they could wear to influence the predicted BMI.

Acknowledgments: The authors wish to thank Glenn Fung for discussions and sharing some of the datasets used in this work, as well as the anonymous reviewers for their comments and suggestions. This work is supported in part by NSF 1545481, 1704117, 1836978, and the MADLab AF Center of Excellence FA9550-18-1-0166.

REFERENCES

[1] Anish Athalye and Ilya Sutskever. 2017. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397* (2017).

[2] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.

[3] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

[4] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*. http://arxiv.org/abs/1412.6572

[5] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[6] Enes Kocabey, Mustafa Camurcu, Ferda Ofli, Yusuf Aytar, Javier Marin, Antonio Torralba, and Ingmar Weber. 2017. Face-to-BMI: Using computer vision to infer body mass index on social media. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*. AAAI press, 572–575.

[7] Enes Kocabey, Ferda Ofli, Javier Marin, Antonio Torralba, and Ingmar Weber. 2018. Using Computer Vision to Study the Effects of BMI on Online Popularity and Weight-Based Homophily. In *Social Informatics*, Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov (Eds.). Springer International Publishing, Cham, 129–138.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[9] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582.

[10] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 582–597. https://doi.org/10.1109/SP.2016.41

[11] World Health Organization. 2018. BMI Classification. http://apps.who.int/bmi/index.jsp?introPage=intro_3.html. (2018). http://apps.who.int/bmi/index.jsp?introPage=intro_3.htm

Figure 3: Attacks forcing BMI predictions into the "normal weight" range [18.7, 24.9]. Row (a): Original BMI prediction $f(X)$. Row (b): Attack δ and its norms. δ's color scale maps [-2, 2] linearly to [0, 255] (gray = no attack). Row (c): Attacked BMI prediction $f(X + \delta)$.

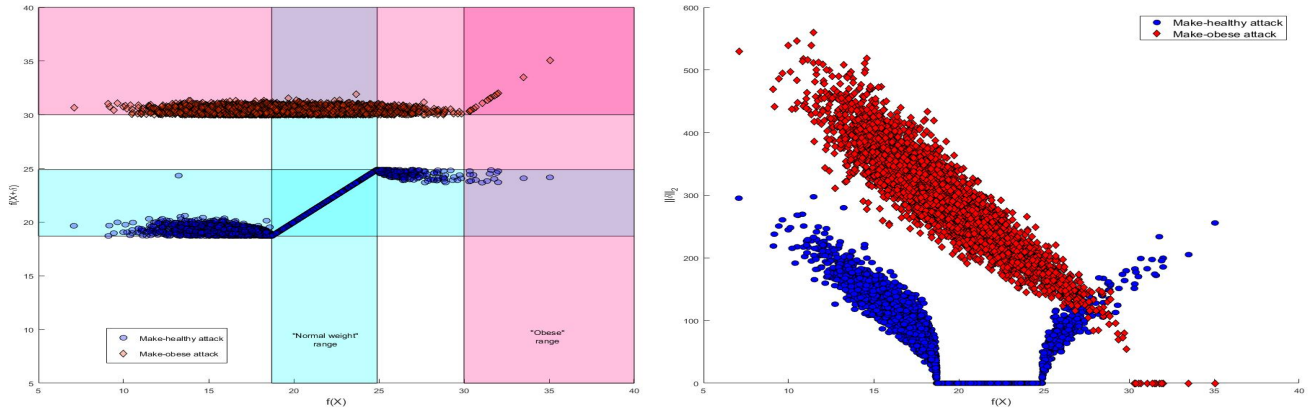

Figure 4: Left: x-axis: the initial BMI prediction $f(X)$, y-axis: the corresponding attacked BMI prediction $f(X+\delta)$ for each image in the VisualBMI data set. We have highlighted the relevant target ranges. Right: x-axis: $f(X)$, y-axis: the corresponding $\|\delta\|_2$ of the first successful rounded δ for each victim image in the VisualBMI data set.

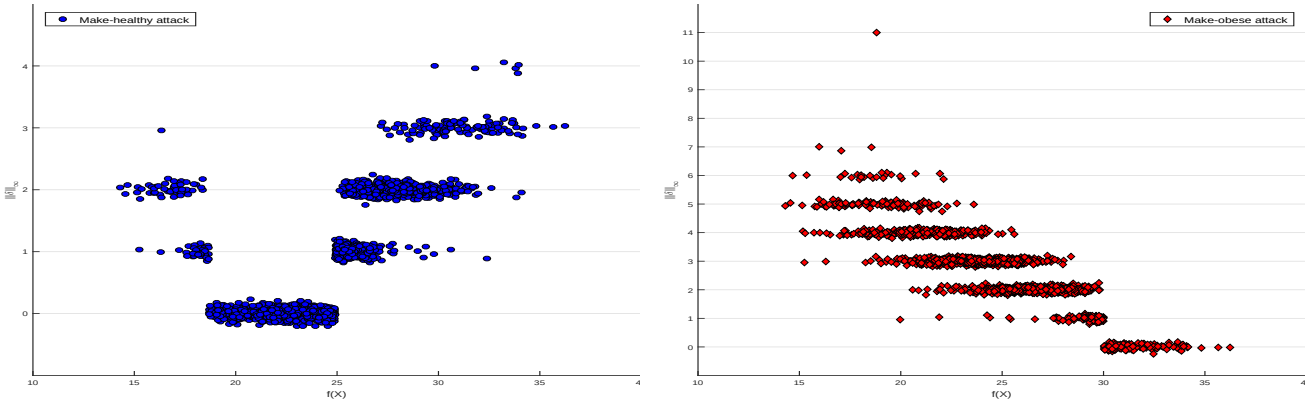

Figure 5: Attack $\|\delta\|_\infty$ on the FCBMI test set for make-healthy (Left) and make-obese (Right) attacks. To help visualize the distribution of data we dithered the norms using iid Gaussian noise with mean 0 and variance .005