Final Project for Deep Learning (CSE331-01)

JungYoon Sung
jypaulsung@khu.ac.kr

Abstract

We tried to perform a black box attack on VGGnet through adversarial image samples created by the source model(ResNet). We started from the baseline model that implements the basic iterative model from [3]. Then, we tried different methods from other papers to achieve the best performance.

1 Single-method approach

1.1 Iterative least-likely class method

First, we tried the iterative least-likely class method described in [3]. By specifying which of the incorrect classes the model should select, we aimed to achieve better performance of the adversarial attack. We expected this method to show some significant difference because our experiment used ImageNet, which has a large number of classes and varying degrees of significance in the difference between classes, as the dataset. However, the prediction accuracy was only reduced by 3%p to 53.25% compared to the baseline model's 56.25%.

2 Hybrid Approach

Since our single-method approach only showed marginal improvement, we tried different hybrid approaches that combined multiple types of FGSM.

2.1 MI-FGSM

We first tried MI-FGSM described in [1]. We expected MI-FGSM, which incorporates momentum into the iterative attack to stabilize the direction of perturbation and escape local minima, to improve transferability. Since our model uses ResNet as the source model and performs an adversarial attack on VGGnet, improving the transferability could enhance the performance. We chose the decay factor of 1.0 as it was proven to be most effective in [1]. It simply adds all previous gradients to perform an update within the iterations. Combining iterative least-likely class method and MI-FGSM showed prediction accuracy of 43%.

2.1.1 TI-FGSM

Our second hybrid two-way approach was incorporating TI-FGSM from [2] to the iterative least-likely class method. TI-FGSM applies a convolutional kernel to average the gradient over a local neighborhood and thus makes adversarial examples translation-invariant. Smoothing the gradients using a convolutional kernel can enhance the attack's performance by making adversarial examples more robust to transformations. We tried with a 3×3 Gaussian Kernel for computational efficiency. This approach reduced the prediction accuracy to 33.25%.

Preprint. Under review.

2.2 DI-FGSM

Our last hybrid two-way approach was combining iterative least-likely class method with DI-FGSM described in [4]. DI-FGSM introduces randomness into input by applying diverse transformations before calculating the gradient. Since this reduces overfitting of perturbations to the source model, it can improve performance when we carry out the adversarial attack on a different model. For transformation operations, we implemented the same method as [4]. We first randomly resized the input images into rnd$\times$rnd$\times$3 image, with rnd in [299,330), and then padded to the size $330 \times 330 \times 3$ in a random manner. The input_diversity function applied this transformation with a probability of 0.5. This approach showed the prediction accuracy of 23.5%.

2.3 MI-FGSM & TI-FGSM

To find out if combining more than one method to the iterative least-likely method could improve the adversarial attack's performance, we tried merging MI-FGSM and TI-FGSM. We used the same 3×3 Gaussian Kernel as our previous implementation and also used momentum of 1.0. We expected this method to show better performance by yielding perturbations that are both more robust and transferable. However, the prediction accuracy actually recovered to 39.38%. It showed better results than adding only MI-FGSM, but still worse than adding only TI-FGSM.

2.4 MI-FGSM & DI-FGSM

Our next attempt was combining MI-FGSM and DI-FGSM to the iterative least-likely method. This approach could make our adversarial examples more generalizable across different models and inputs. We first applied the same input_diversity function with probability of 0.5 and used a momentum of 1.0. This actually showed some significant improvement as its prediction accuracy dropped to 16.25%.

2.5 TI-FGSM & DI-FGSM

We also tried merging TI-FGSM and DI-FGSM into the iterative least-likely method. By combining translation invariance and input diversity, we expected the adversarial samples to evade transformations and preprocessing-based defense in the VGGnet. This time we used the same input_diversity function with the same probability for DI-FGSM. However, we tried to improve the performance with a 5×5 Gaussian kernel for the TI-FGSM. This approach yielded a prediction accuracy of 23.63%.

2.6 Combining all MI-FGSM & TI-FGSM & DI-FGSM

Since our previous hybrid approaches proved that combining more than one FGSM could result in better performance of the adversarial attack, we decided to merge all three methods with the iterative least-likely method. A 5×5 Gaussian kernel, same input_diversity function, and momentum of 1.0 were chosen. The prediction accuracy was 17.75%, the best so far.

Table 1: Performance Comparison of Different Adversarial Attack Methods

Group	Method	Prediction Accuracy (%)
Single-method	Baseline	56.25
	L.L. Class	**53.25**
Hybrid 2-way approach	L.L. Class with MI-FGSM	43.00
	L.L. Class with TI-FGSM	33.25
	L.L. Class with DI-FGSM	**23.50**
Hybrid 3-way approach	L.L. Class with MI-TI-FGSM	39.38
	L.L. Class with MI-DI-FGSM	**16.25**
	L.L. Class with TI-DI-FGSM	17.75
	L.L. Class with MI-TI-DI-FGSM	**16.13**

3 Final Version

To fine-tune our hybrid approach, we experimented with Gaussian kernels of different sizes. Our final version uses a 15×15 Gaussian kernel to implement TI-FGSM and showed the prediction accuracy of 16.13%. Making the Gaussian kernel larger did not yield significant improvements considering its computational trade-off. The results shown in Table 1 indicate that DI-FGSM was the most effective method for creating adversarial samples against VGGnet with ResNet as the source model. DI-FGSM's advantage over MI-FGSM and TI-FGSM can be addressed with 2 explanations in our understanding. The MI-FGSM may have resulted in perturbations overfitted to the source model due to momentum accumulation. Additionally, for TI-FGSM, translation-invariance may not have been sufficient to account for differences in how ResNet and VGGNet process input features.

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018.

[2] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4312–4321, 2019.

[3] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[4] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2019.