This paper talks about zeroth order optimization-based attacks on deep neural networks without training the substitute model. Deep neural networks (DNNs) are used to perform image classification, text mining, speech processing, etc. There are rising concerns about the robustness of DNNs for tasks like traffic sign identification for autonomous driving. These tasks require a high level of security, as it is easy for someone to include a set of images that might lead to misclassification. The paper gives us a brief view of the existing attacks such as the Fast Gradient Sign Method, Jacobian-based saliency Map Attack, DeepFool, and Carlini & Wagner (C&W) Attack. An important issue of transferability is addressed. Adversarial examples that are misclassified by one model are likely to be misclassified by another model as well - Transferability raises concerns for security as the above methods prove that it is easy to craft malicious images even if the internal configurations of DNN are unknown. The author also gives us the difference between black box and white box attacks with the major difference being no knowledge of the internal DNN structure in a black box Attack. A black box attack is most effective with a substitute model which is placed above the targeted model, on this substitute model, the user has full access to the internal configurations, which he can change iteratively to behave like the targeted DNN. ZOO the proposed form of attack uses zeroth-order methods. The proposed black-box attack in this paper leverages the characteristics of ZOO to avoid the creation and training of a substitute model. ZOO must be used with the correct optimization method, like a stochastic gradient method, especially for large image inputs. ZOO, like SGD, iteratively optimizes and updates a coordinate or a batch of coordinates instead of computing full gradients for efficient updates.

Algorithm 1 Stochastic Coordinate Descent

1: **while** not converged **do**
2: Randomly pick a coordinate $i \in \{1, \ldots, p\}$
3: Compute an update δ^* by approximately minimizing

$$\arg\min_{\delta} f(\mathbf{x} + \delta \mathbf{e}_i)$$

4: Update $\mathbf{x}_i \leftarrow \mathbf{x}_i + \delta^*$
5: **end while**

At each iteration, a coordinate is selected at random and updated by approximately minimizing the objective function along that coordinate and the best δ is decided based on the estimated gradient and Hessian for the image using any 1st or 2nd method. For this, ADAM and Newton are the two methods used in which ADAM proves to be the better method experimentally.

We are then presented with some experimental results and performance analysis on the MNIST and CIFAR-10 datasets which result in ZOO achieving a nearly 100% success rate. The L2 distortion rates are close to the ones in C&W white-box attack and conclude that the success rate for ZOO is higher than substitute-based attack methods, especially for targeted attacks. On further experimentation by using hierarchical attack, and importance sampling a clear distinction in performance is provided.

Table 3: Comparison of different attack techniques. "First Valid" indicates the iteration number where the first successful attack was found during the optimization process.

Black-box (ZOO-ADAM)	Success?	First Valid	Final L_2	Final Loss
All techniques	Yes	15,227	3.425	11.735
No Hierarchical Attack	No	-	-	62.439
No importance sampling	Yes	17,403	3.63486	13.216
No ADAM state reset	Yes	15,227	3.47935	12.111

We can conclude from the entire paper that ZOO is a new type of attack on DNN without needing a substitute model which gives comparable results to C&W's white box attacks and significantly outperforms the substitute-model-based attacks.

References: *Chen, Pin-Yu, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. ACM workshop on artificial intelligence and security, 2017.*