**Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning**

**Problem Statement**

The article entitled “Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning” by Chen et al. addresses the problem of backdoor attacks on deep learning systems, specifically focusing on backdoor poisoning attacks. Backdoor attacks aim to create a backdoor in a learning-based authentication system, allowing an attacker to easily circumvent the system by leveraging the backdoor. The goal is to mislead the victim learning system into classifying certain instances as a target label specified by the attacker. The researchers aim is to investigate and demonstrate the feasibility of backdoor poisoning attacks under a weak threat model, where the attacker has no knowledge of the model or training set used by the victim system, is allowed to inject only a small amount of poisoning samples, and the backdoor key is difficult to notice even by humans.

**Summary**

The researchers introduce backdoor attacks on deep learning systems, specifically backdoor poisoning attacks, where the attacker injects poisoning samples into the training set to create a backdoor in a learning-based authentication system. They emphasize a weak threat model, where the attacker has no knowledge of the system's model or training set, can only inject a small number of poisoning samples, and the backdoor remains stealthy. The study demonstrates the feasibility of these attacks and shows that an attacker can achieve an attack success rate of over 90% by injecting approximately 50 poisoning samples.

**Critical Points - Strong and Weak Points**

**Strong Points:**

1. Feasibility demonstration: The researchers provide evidence of the feasibility of backdoor poisoning attacks under a weak threat model, highlighting the potential threats to deep learning systems.
2. Small number of poisoning samples: They show that a successful attack can be achieved with a relatively small number of poisoning samples (around 50), making the attack practical for adversaries.
3. Physically implementable backdoors: The concept of creating backdoors without touching the training process adds to the realism and practicality of the attacks.

**Weak Points:**

1. The evaluation mainly focuses on the attack success rate and does not extensively cover the robustness of the proposed attack strategies against potential defense mechanisms.
2. Although defense strategies are briefly explored, the research does not propose effective defense mechanisms against backdoor poisoning attacks.

**Missing Technology**

The article does not mention any specific technologies. This article could benefit from the exploration of advanced defense mechanisms or the countermeasures specifically designed to detect and mitigate backdoor poisoning attacks in deep learning systems.

**Proposing a Better Solution**

To improve the research, it would be beneficial to develop effective defense strategies against backdoor poisoning attacks. This could involve by exploring effective detection techniques, model introspection methods, or adversarial training approaches to identify and mitigate the presence of poisoning samples or backdoor instances. Furthermore, researching secure training protocols or robust learning algorithms that can resist poisoning attacks could provide stronger protection against backdoor threats in deep learning systems. In my opinion interpretability in deep learning models may help to uncover the presence of hidden backdoors.

**References**

Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.