CSE3000 Research project

Imperceptibility of the WaNets backdoor attack and its impact on deep regression models

J.G.C. van de Meene

May 8, 2024

Introduction

In recent history, development and usage of deep learning has exploded- and for good reason. Applications of this technology have proven it to be incredibly valuable across many fields. The trade-off for the capabilites deep learning models provide is the resources required to train them. As such, pretrained models have become popular too, but this now common practice has been shown to come with some serious security risks. During training or fine tuning these models can be poisoned that cause them to function normally unless a certain trigger patterns appear. What patterns it triggers and how it affects the functionality depends on how the attacker designs the backdoor model, who very well may have malicious intents.

This threat seems to be indiscriminate in regards to the task at hand, as it has been researched in the context of image recognition (Chen et al., 2017), speech recognition (Liu et al., 2018), language processing (Dai et al., 2019), or reinforcement learning (Commission et al., 2020). Many of the attack designed for this research are still imperceptible as they distort the data in a way easily detectable by humans, which is a limitation that has been designed to overcome by the WaNets backdoor attack (Nguyen and Tran, 2021).

It is clear that backdoor attacks on deep learning models have been subject to research for a while now- rightfully so. Although the risk is evident within a wide range of tasks, most research is still focussed on certain branch of deep learning utilizing deep classification models. Not all applications of this technology are designed to assign classes to data, however. Many problems require predictions to be made in a less discrete manner. An example of such problem is the task of gaze estimation where the direction a person is looking at is estimated. A direction is usually a certain vector in a continuous solution space, representing the orientation of the gaze within a three-dimensional coordinate system. In such cases deep regression models are more applicable, but whether or not they are as vulnerable against backdoor attacks is not yet entirely clear. This leaves a much less-explored, but equally important area of research to be investigated. Asside from whether attacks affect regression models the same way they affects classification models, the imperceptability of the attacks may also differ from the latter.

This research aims to answer among others the aforementioned uncertainties. In particular the WaNets backdoor attack will be applied to a Gaze estimation model and compared to the existing studies on deep classification models in regards to the vulnerability of the model and the imperceptibility of the attack.

References

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

Commission, E., Centre, J. R., Sanchez, I., Junklewitz, H., & Hamon, R. (2020). *Robustness and explainability of artificial intelligence â from technical to policy solutions*. Publications Office. https://doi.org/doi/10.2760/57493

Dai, J., Chen, C., & Li, Y. (2019). A backdoor attack against lstm-based text classification systems. *IEEE Access*, *7*, 138872–138878. https://doi.org/10.1109/ACCESS.2019.2941376

Liu, Y., Shiqing, M., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., & Zhang, X. (2018). Trojaning attack on neural networks. https://doi.org/10.14722/ndss.2018.23300

Nguyen, T. A., & Tran, A. T. (2021). Wanet - imperceptible warping-based backdoor attack. *International Conference on Learning Representations*. https://openreview.net/forum?id=eEn8KTtJOx