Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

SECURITY AND PRIVACY

ASSIGNMENT -1

Submitted By

Harsh Dhingra

19323904

1. Topic Abstract

The dissertation is being done under the guidance of Dr. Gerald Lacey (*Associate Professor in Graphics & Vision*). The topic of my dissertation is **ASSESSING AND RANKING OF SURGICAL SKILLS** using a pairwise deep ranking model. Surgical Skills are still being accessed manually; moreover, most surgical training programs evaluate their surgical trainees on a subjective approach. Surgical training and assessment have been chastised in the past due to lack of credibility; according to a case study, surgical errors are one of the five most adverse events in operation, and the latest trends state that surgical mistakes happen more than 4000 times a year just in the United States. Much research for assessing surgical skills is being carried on to shift it from a manual to an automated approach. This would not only reduce the burden on the examiner, but this would bring generalization and universality all over the globe.

Due to the increase in the number of surgeons over the past decade and the ongoing pandemic has the reduced face-to-face interaction between tutor and the trainee; this has eventually forced a trainee to seek help from the web. The web consists of hundreds of thousands of how-to-tutorials be it suturing, knot tying, needle-passing, etc. - this can confuse a learner as to which technique to be chosen which is the best technique, taking account these factors, there is a need for classifying which method is better and which is best so that one can learn the best approach and could also test their technique. The research aims to build a deep learning model that can rank the videos pairwise on a publically available dataset; this model future can be used by users who are presumably doctors to access the pre-existing tutorial to learn and to access their skills too,

Since such a system would have a broad user base and as the model expands, more data would be collected along that therefore I need to consider security and privacy concerns

2. Security Considerations

Deep learning models are of great use and are being widely adopted these days and are used in many applications in life relating to predictions and classification. The model that I am developing has significant application in the healthcare field; therefore, any false prediction or errors could do great harm; thus, identifying attacks on the system becomes very important.

Various studies on the possible attacks that any deep learning model could face; concluded that attack scenarios differ by the amount of information the attacker has about the model.

Different types of attacks possible on my project –

- ✓ Evasion Attacks – Designing an input in such a way as to get wrong results from the model, i.e., the network is fed with adversarial examples. The attacker can design the input in such a way that the model gives incorrect output, and this would highly affect the model.

- ✓ Poisoning Attacks – This is an attack on a deep learning model where the attacker adds examples to the training set or modifies the training set in order to manipulate the behavior of the model; this could be very hazardous to the model.[6]

Defense Techniques against potential attacks

- ✓ Adversarial training: defense method against adversarial samples, it improves the robustness of a neural network by training it with adversarial examples [1]

- ✓ Gradient Masking: Using gradient-based defense algorithm [6]

- ✓ Randomization :[3] According to various researches, randomization has proven to show good results, several typical randomization techniques are
    - Random input transformation – which is random resizing and padding at the interference time
    - Random Noising – by adding a random noise layer before each convolutional layer in both training and testing.

Several other techniques to mitigate attacks include GAN-based input cleansing (A defense GAN algorithm) , Feature denoising, and auto-encoder-based input denoising.[3].

3 Privacy Considerations

3.1 IMPLICIT CONSIDERATIONS

Privacy is a crucial consideration in any deep learning model; A deep learning model should take special care when dealing with training, testing, and validation dataset. It should be prioritized that there should be no violation of user privacy.

In the final year project, I am using the publicly available dataset published by JIGSAWS(JIGSAWS: The JHU-ISI Gesture and Skill Assessment Working Set) there is no user-sensitive data present in the dataset. Also, if I needed more data, I would be reaching out to some private data repositories, and this would be done after formal approval from the college ethics committee.

3.2 EXPLICIT CONSIDERATIONS

Around a deep learning model, many privacy concerns revolve, majorly they are related to sensitive input data either during training or inference; therefore, the skill accessing system should not leak any information apart from what can be deduced from an accessing system, that is the output should not help the attacker in any possible way so that he or she can identify any sensitive information (any personal information or any other attributes).

An ideal skill accessing system should consider providing proper privacy protection while giving good accuracy, but always there is a compromise between privacy or efficacy of the system.

Various researches are being carried out so that there is no compromise between the efficacy and the privacy of any deep learning model, some of the privacy-preserving techniques that I could be exploring are as follows

- ✓ Homomorphic encryption: a public key cryptographic scheme, it's a type of encryption that preserves operational integrity; basically, all operations are done on ciphertexts. Homomorphic encryption creates ciphers in such a way that there is a parallel operation that you could run on the cipher data. [5][2]

- ✓ Garbled Circuits: provides a mechanism for building two secure parties to evaluate an arbitrary function without leaking the information regarding input irrespective of the output.[2]

- ✓ Split learning: This is new collaborative learning; it aims to protect user data privacy without revealing the input data to the server. It simultaneously runs DNN's where the model is split into two parts, one for the client and the other at the server. Its very communication efficient[4][5]

I would go through all general security practices like choosing strong encryption (for databases) and go through Europe's new data and security law GDPR for security and privacy consideration while designing my skill accessing the system.

References

[1] https://www.sciencedirect.com/science/article/pii/S209580991930503X

[2]https://www.hindawi.com/journals/misy/2020/6535834/

[3]Samangouei P, Kabkab M, Chellappa R. Defense-GAN: protecting classifiers against adversarial attacks using generative models. 2018. arXiv:1805.06605. https://www.sciencedirect.com/science/article/pii/S209580991930503X#bb0395

[4]]https://dl.acm.org/doi/10.1145/3320269.3384740#:~:text=A%20new%20collaborative%20learning%2C%20called,the%20other%20for%20the%20server.

[5]https://www.datacouncil.ai/talks/data-security-and-privacy-in-the-age-of-machine-learning

[6] https://arxiv.org/pdf/1807.11655.pdf

[7]https://link.springer.com/content/pdf/10.1007/s42979-020-00254-4.pdf