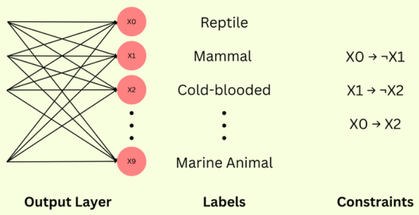
# TOWARDS BENCHMARKING THE ROBUSTNESS OF NEURO-SYMBOLIC LEARNING AGAINST BACKDOOR ATTACKS

**Diego Becerra Merodio**
**d.becerramerodio@student.tudelft.nl**
Supervisor: **Andrea Agiollo** Professor: **Kaitai Liang**

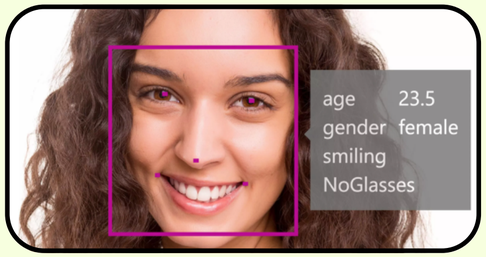
## 1 Background

**Semantic Loss** is a type of NeSy model that allows a user to define constraints for the output of a model.




**BadNets** is a type of backdoor attack where the attacker alters the images by adding a digital modification that is consistently found on the poisoned samples. The label(s) of these samples are also modified to contain the target label.

**CelebA** is a dataset that consists of 202,599 faces, which are labeled with 40 attributes.



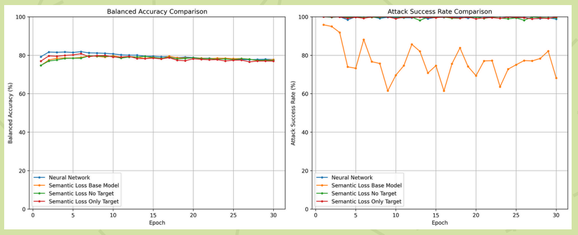

## 2 Research Question

How do semantic loss models perform against BadNets data poisoning attacks?

## 3 Model Comparisson

Evaluated four different models:

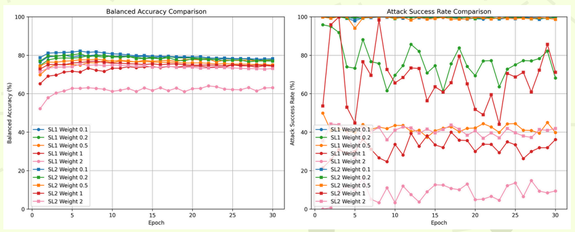
- Neural Network
- SL with Base constraints
- SL with Targetless constraints
- SL with Target-Focused constraints




- Base → More robust
- Targetless → No improvement
- Target-Focused → SL component too small

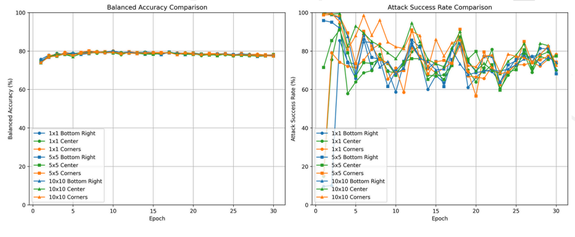
## 4 Different Weights

Evaluated Base and Target-Focused models with a SL component of weight 0.1, 0.2, 0.5, 1, and 2.




- Small weight → High BA and ASR
- Large weight → Decrease in BA and ASR
- Important to tune weight for optimal results

## 5 Trigger Exploration




Evaluated nine different triggers by combining the sizes 1x1, 5x5, and 10x10 with positions bottom right corner, center, and all corners + center.

No effect on the performance of the attack.

## 6 Limitations

- Computational resources
- Third-party semantic loss library
- CelebA dataset label inconsistency and imbalance

## 7 Conclusion

- SL increase robustness of NNs
- Constraints play a very significant role
- Higher SL weights improved robustness but reduced BA.
- Trigger size and position had minimal impact on ASR or BA.

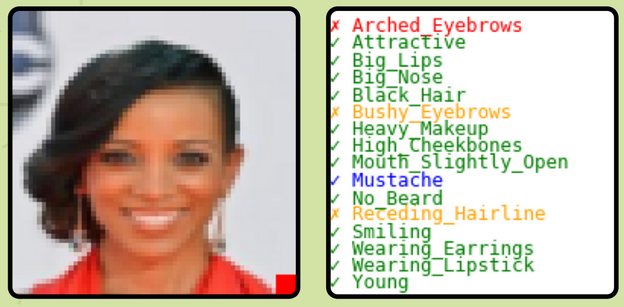



Fig.Sample image with corresponding labels.

## References

[1] A. d'Avila Garcez and L. C. Lamb, "Neurosymbolic ai: the 3rd wave," Springer, 2023. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10462-023-10448-w.pdf
[2] B. P. Bhuyan, A. Ramdane-Cherif, R. Tomar, and T. P. Singh, "Neuro-symbolic artificial intelligence: a survey," Springer, 2024. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s00521-024-09960-z.pdf
[3] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," IEEE, vol. 35, no. 1, 2024. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9802938
[4] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," 2020. [Online]. Available: https://arxiv.org/pdf/2007.10760
[5] D. Kahneman, Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.
[6] H. A. Kautz, "The third ai summer: Aaai robert s. engelmore memorial lecture," AI Magazine, vol. 43, no. 1, pp. 105–125, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12036
[7] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. V. den Broeck, "A semantic loss function for deep learning with symbolic knowledge," Proceedings of the 35th International Conference on Machine Learning, 2018. [Online]. Available: https://proceedings.mlr.press/v80/xu18h/xu18h.pdf
[8] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," IEEE, vol. 7, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8685687
[9] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in Proceedings of International Conference on Computer Vision (ICCV), December 2015.
[10] R. Torfason, E. Agustsson, R. Rothe, and R. Timofte, "From face images and attributes to attributes," 11 2016.