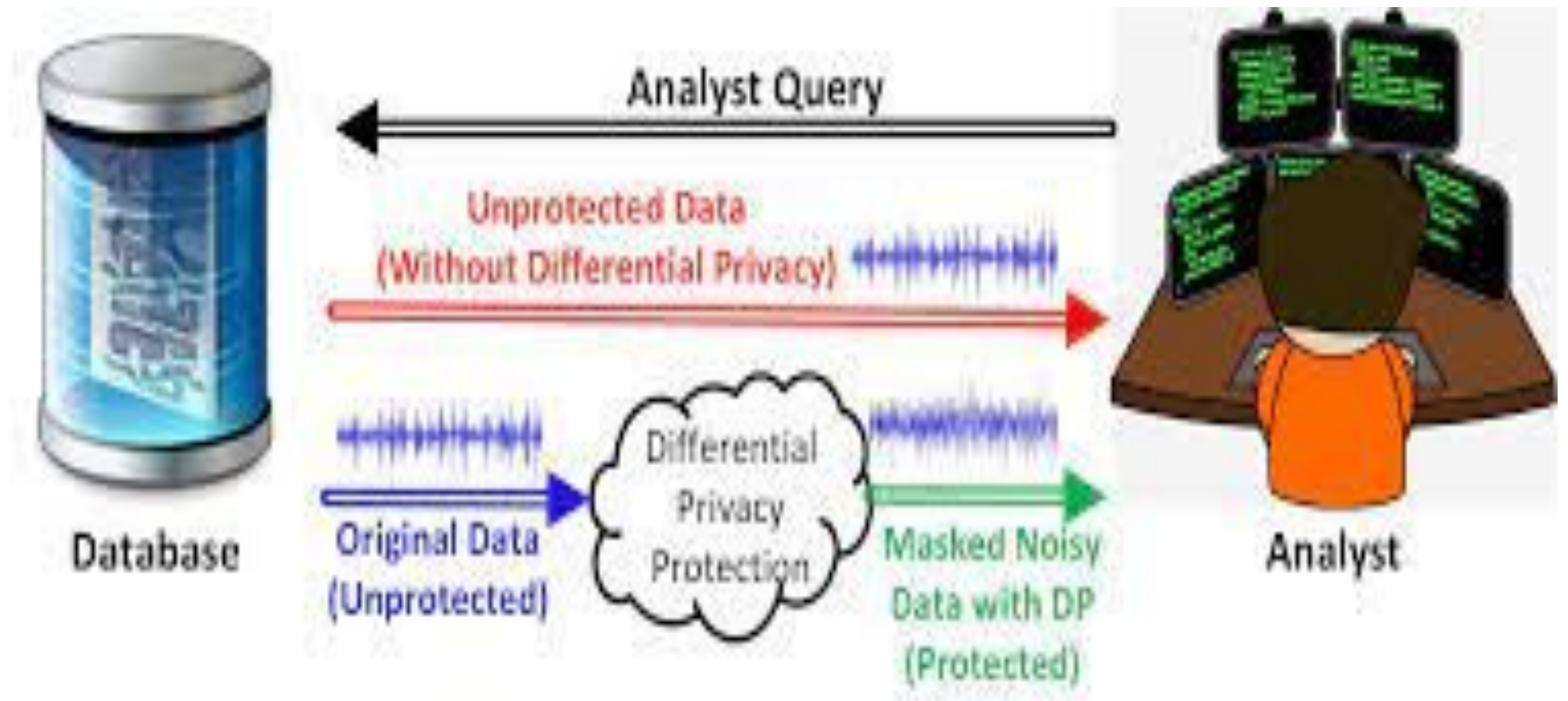
When Differential Privacy Meets Interpretability: A Case Study

Aman Priyanshu*[1], Rakshit Naidu*[1,2,4], Aadith Kumar[1,3],
Sasikanth Kotti[4,6], Haofan Wang[2,4], Fatemehsadat Mireshghallah[4,5]
[1]Manipal Institute of Technology, [2]Carnegie Mellon University, [3]University of Pennsylvania,
[4]OpenMined, [5]University of California, San Diego, [6]IIT Jodhpur

CVPR VIRTUAL JUNE 19-25

INTRODUCTION

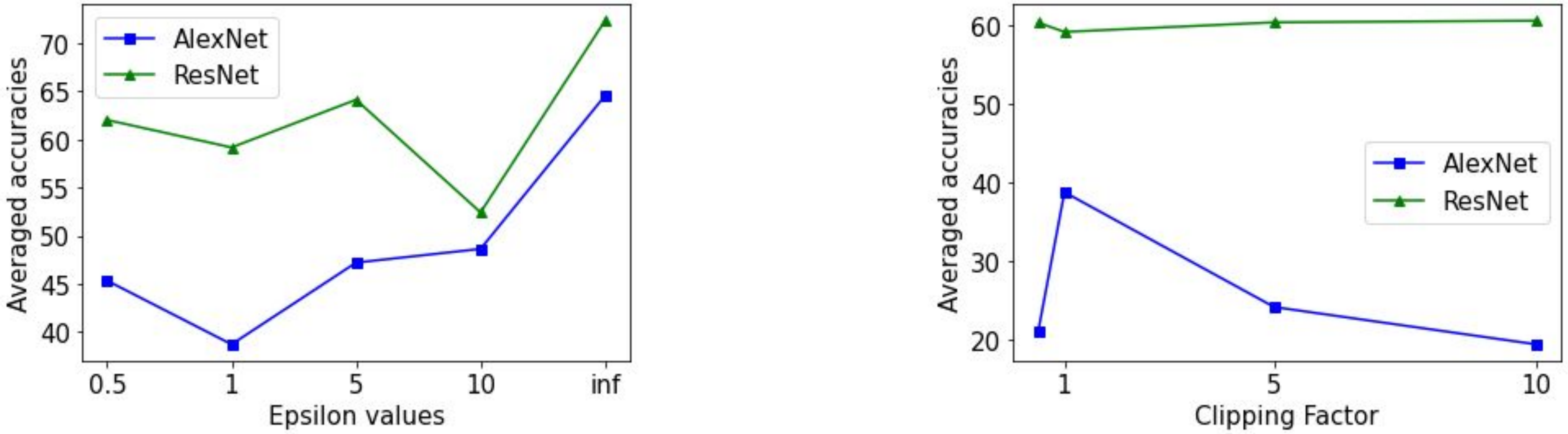
- Differential Privacy (DP) is an extensive tool that constitutes strong privacy guarantees for algorithms on a given dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset.



- We propose benchmarking DP-DNNs of different sizes, trained with different levels of privacy (ϵ) and evaluating their interpretability, to gain a better insight into how privacy plays into model interpretability.

- We utilize Grad-CAM as our interpretability method and use APTOS (a real-world medical dataset) to train our models. The noise used is sampled from a Gaussian Distribution $G(0,\sigma^2)$

PRELIMINARY EVALUATIONS



Averaged accuracies of inputs masked with their explanations over AlexNet and ResNet networks for (a) different ε values and (b) different clipping factors (S) with ε = 1. We quantitatively show that there's a significant gap between privacy and explanation quality.

CONCLUSION

- DP-trained models allow more flexibility than standard models due to the privacy factor (ε) and could be commercially viable for medical imaging tasks.

- As future work, we hope to understand the effect of visual explanations by adding Local DP (noise at the data level; rather than at the training level) and we also wish to release a framework for Interpretability specifically catered to DP-trained models.

REFERENCES

- Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Oct 2016.

- Sahib Singh and Harshvardhan Sikka. Benchmarking differentially private residual networks for medical imagery. CoRR, abs/2005.13099, 2020.