CNNs VS imbalanced datasets

1. Background

- Convolutional Neural Networks (CNN) widely used[1]
- Mostly viewed as Black Boxes[1]
- What are Imbalanced dataset?[2]
- Imbalance is not uncommon[3]
  - Harder to get samples of rare diseases

2. How do imbalanced training datasets affect the performance of CNNs?

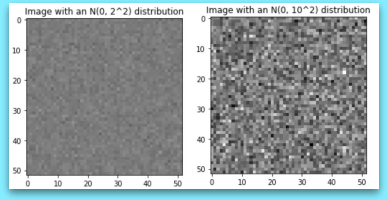
- Performance of network trained on:
  - Balanced datasets
  - Datasets with missing targets
  - Dataset with normally distributed targets
- Related work shows:
  - networks trained on balanced datasets significantly outperform others[3-10]

3. CNN

- Shallow network[11]
- Adam optimizer[12]
- Standard deviation

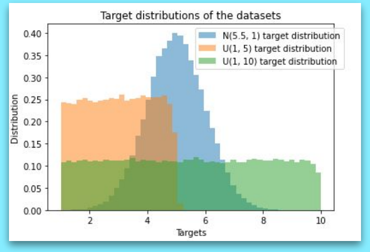
4. Datasamples

- Samples have $N(0, x^2)$ distribution
  - x is drawn from target distribution dependant on dataset
- Visualisation of $N(0, 2^2)$ and $N(0, 10^2)$ samples:



5. Datasets

- Synthetic
- U(1, 10)
  - reference dataset
- U(1, 5)
  - missing targets
- N(5.5, 1)
  - different distribution



Ratish Thakoersing

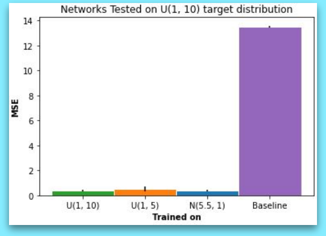
Delft University of Technology

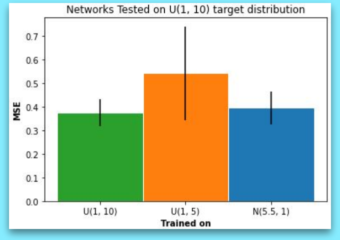
CSE3000

TU Delft

6. Results

- Mean-squared error
- Baseline
- All networks tested on U(1, 10) datasets
- Experiments repeated 10 times





7. Discussion and conclusion

- All networks significantly outperformed the baseline
  - Networks were able to learn the task with imbalanced datasets
- The networks trained on the balanced datasets had the best performance
  - In line with hypothesis
- The networks trained on the datasets with normally distributed targets performed slightly worse
  - Some targets underrepresented -> networks were not able to predict those as effectively
- The networks trained on the datasets with missing targets had the worst performance
  - Training sets did not include all of the targets of the test sets -> the networks were unable to perform as well as the networks trained on other target distributions

References

1. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning.nature, 521(7553):436–444, 2015.
2. Paula Branco, Lu´ıs Torgo, and Rita Ribeiro. Smogn: a pre-processing approach for imbalanced regression. 092017.
3. Aida Ali, Siti Mariyam Shamsuddin, and Anca LRalescu. Classification with class imbalance problem.Int. J. Advance Soft Compu. Appl, 5(3), 2013.
4. Paulina Hensman and David Masko. The impact of imbalanced training data for convolutional neural net-works. Degree Project in Computer Science, KTH Royal Institute of Technology, 2015.
5. Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),June 2016.
6. Claudia Beleites, Richard Baumgartner, Christopher Bowman, Ray Somorjai, Gerald Steiner, Reiner Salzer,and Michael G Sowa. Variance reduction in estimating classification error using sparse datasets.Chemometrics and intelligent laboratory systems, 79(1-2):91–100,2005.
7. Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. CoRR, abs/2102.09554, 2021.
8. Paula Branco, Lu´ıs Torgo, and Rita Ribeiro. Smogn: a pre-processing approach for imbalanced regression. 092017.
9. Paula Branco, Lu´ıs Torgo, and Rita Ribeiro. Rebagg: Resampled bagging for imbalanced regression. 09 2018.
10. Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4):221–232, 2016.
11. Daniel E. Kim and Mikhail Gofman. Comparison of shallow and deep neural networks for network intrusion detection. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC),pages 204–208, 2018.
12. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.