

Bringing the Discussion of Minima Sharpness to the Audio Domain: a Filter-Normalised Evaluation for Acoustic Scene Classification

Manuel Milling^{1,2}, **Andreas Triantafyllopoulos**^{1,2}, **Iosif Tsangko**²,
Simon D. N. Rampp², **Björn W. Schuller**^{1,2,3}

- 1: Chair of Health Informatics,
Technical University of Munich, MRI, Germany
- 2: Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
- 3: Group on Language, Audio & Music,
Imperial College London, UK

ICASSP 2024, Seoul, April 17, 2024



Motivation

- Despite good performance, training and generalisation of artificial neural networks (ANNs) only partially understood
- Common evaluations only concerned with development performance
 - Particular challenges of expressiveness for out-of-distribution (OOD) data
- Sharpness (or flatness): how fast the loss function changes when varying parameters in a minimum [1]
- Flat minima often considered desirable for generalisation
 - Implications for optimiser design [2] or model selection

[1] Sepp Hochreiter and Jürgen Schmidhuber, “Flat minima,” *Neural computation*, vol. 9, no. 1, pp. 1–42, 1997.

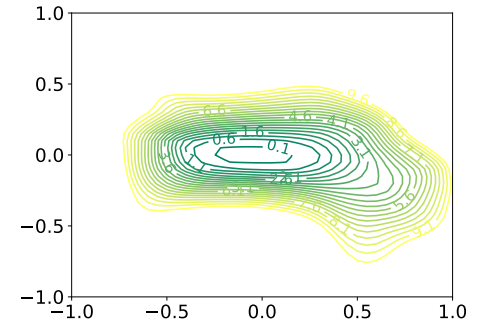
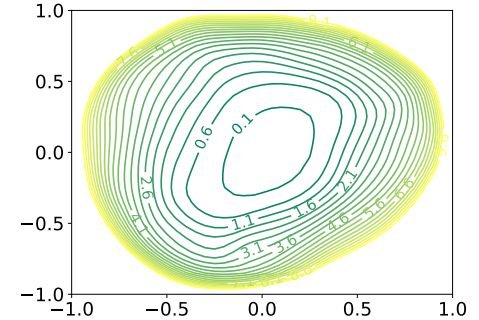
[2] Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2020). SSharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Motivation

- Discussion mostly limited to
 - In-domain (ID)
 - Computer vision (CV)
 - Benchmark datasets (CIFAR10, ...)
 - Sometimes artificial training settings
- Our goal: investigating sharpness in practical training settings for Acoustic scene classification (ASC)
 - Robustness
 - Correlation with generalisation (test accuracy)
 - OOD data vs ID data
 - Effects of different hyperparameters

Sharpness

- Assumption: an ANN with parameters θ (to some degree) converged to a local minimum θ^*
- Sharpness (and flatness respectively) refer to how quickly the loss function changes when moving away from the minimum



Sharpness

- Closely connected to Hessian matrix (curvature)
→ Computationally very expensive
- No unified definition/approach for calculation
- Mostly computed based on differences in the loss function in (random) directions, e.g. in 2D: (away from the minimum)

$$f(\alpha, \beta) = L(\theta^* + \alpha\delta + \beta\eta).$$

L : loss function

δ, η : random directions with dimensions equal to parameters θ

Filter-Normalisation

- Filter-normalisation introduced to counteract different scales of parameters [3]

$$\delta_{i,j} \leftarrow \frac{\delta_{i,j}}{\|\delta_{i,j}\|} \|\theta_{i,j}\|$$

$\delta_{i,j}$, $\theta_{i,j}$: components of the j th filter of the i th layer of the random directiona and parameters, respectively

$\|\cdot\|$: Frobenius norm

- Led to impactful discussion on drivers of flat minima and their benefits for generalisation in loss landscape visualisations
- No quantitative evaluation performed [3]

[3] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein, “Visualizing the loss landscape of neural nets,” in NIPS’18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., 2018, pp. 6391–6401.

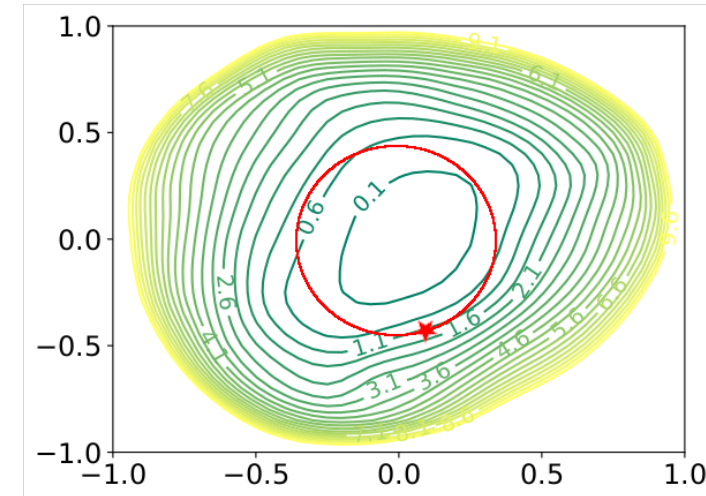
ε -Sharpness

- For quantification we rely on popular ε -sharpness [4]

$$s_\varepsilon = \frac{\max_{\theta \in B(\varepsilon, \theta^*)} (L(\theta) - L(\theta^*))}{1 + L(\theta^*)} \times 100,$$

$B(\varepsilon, \theta^*)$: (High-dimensional) ball of radius ε

- in 2D equals to point with highest loss within a circle around θ^*



[4] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in International Conference on Learning Representations, 2017.

Experiments

- DCASE2020-task 1 acoustic scene classification dataset [5]
 - 10 s audio recordings (64 h total)
 - 10 different acoustic scenes recorded in 10 European cities
 - 3 real recording devices, 6 simulated devices
- PANNs models CNN10 and CNN14 (without pertaining) [6]
- Common optimisers: SGD (with momentum), Adam
- Non-standard optimisers: gradient descent: the ultimate optimiser (GDTUO), Kronecker-factored approximate curvature (KFAC)
- Best development performance after a maximum of 50 epochs

[5] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, “Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), 2020, pp. 56–60.

[6] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020.

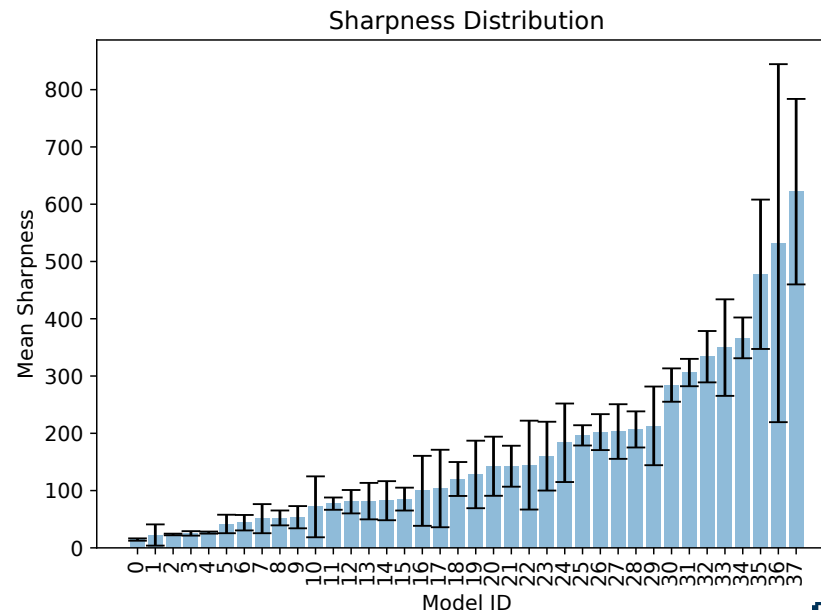
Experiments

- Excluding non-converged models: 38 trained model states

Network	CNN10, CNN14
Optimiser	SGD, Adam, GDTUO, KFAC
Learning Rate	10^{-3} , 10^{-4} , 10^{-5}
Batch Size	16, 32
Random Seeds	42, 43

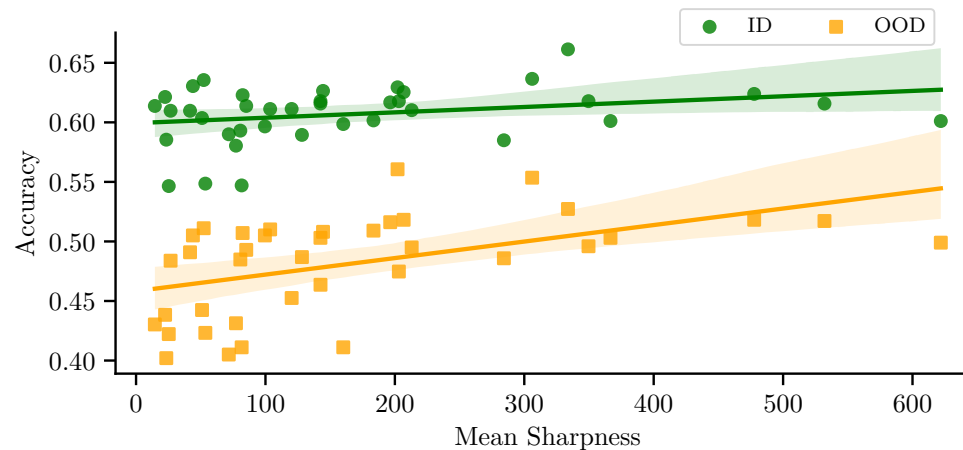
Robustness

- 3 2D-sharpness values with different random directions
- Mean sharpness and standard deviation are reported
- Overall reasonable degree of robustness
- Some models with high deviations (e.g. ID 36)



Generalisation

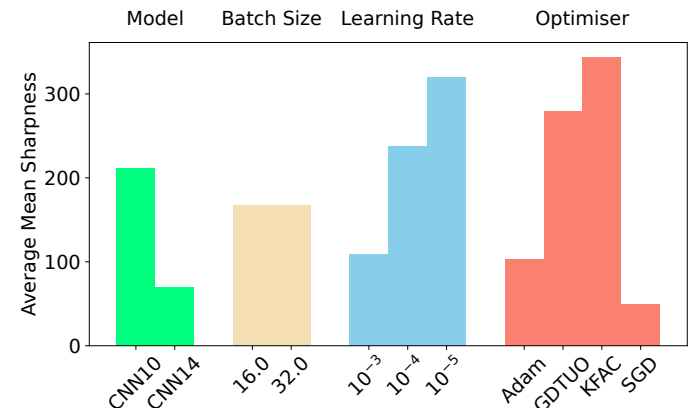
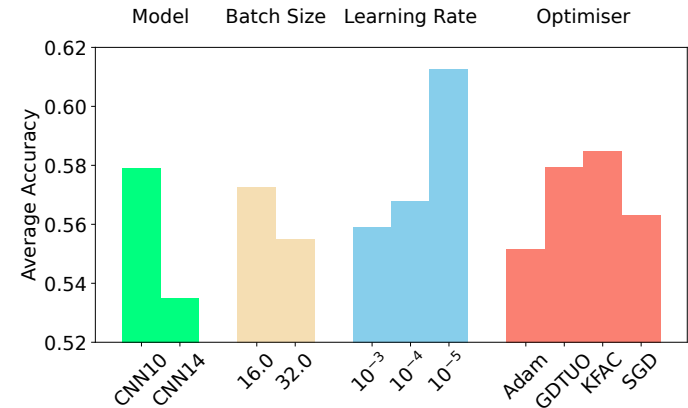
- Positive correlation between sharpness and generalisation
- Effect even stronger for OOD data with generally lower performance
→ OOD = devices (microphones) not present in training data
- Surprising finding: few studies with positive correlation [7]



[7] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion, “A modern look at the relationship between sharpness and generalization,” arXiv preprint arXiv:2302.07011, 2023.

Impact of Hyperparameters

- Similar impacts of hyperparameters on accuracy and sharpness
- Optimisers with highest difference in impact



Limitations

- Only one dataset and one sharpness measure explored
- Some limitations through robustness in sharpness measure
- Convergence-state of models not further considered

Conclusions

- Analysis of filter-normalised 2D ε -sharpness under common training conditions for ASC tasks
- Reasonable robustness across random directions
- Sharper minima correlate with better generalisation (in this case)
→ Not necessarily all insights from CV apply to CA
- Optimisers having strong impact on sharpness
- Further evaluations on audio data necessary for a better picture
- Code, trained model states and visualisations available:
https://github.com/EIHW/ASC_Sharpness

References

- [1] Sepp Hochreiter and Jürgen Schmidhuber, “Flat minima,” *Neural computation*, vol. 9, no. 1, pp. 1–42, 1997.
- [2] Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2020). SSharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412.
- [3] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein, “Visualizing the loss landscape of neural nets,” in *NIPS’18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2018, pp. 6391–6401.
- [4] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang, “On large-batch training for deep learning: Generalization gap and sharp minima,” in *International Conference on Learning Representations*, 2017.
- [5] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, “Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60.
- [6] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [7] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion, “A modern look at the relationship between sharpness and generalization,” arXiv preprint arXiv:2302.07011, 2023.