

# BRINGING THE DISCUSSION OF MINIMA SHARPNESS TO THE AUDIO DOMAIN: A FILTER-NORMALISED EVALUATION FOR ACOUSTIC SCENE CLASSIFICATION

Manuel Milling<sup>1,2</sup>, Andreas Triantafyllopoulos<sup>1,2</sup>, Iosif Tsangko<sup>2</sup>,  
Simon David Noel Rampp<sup>2</sup>, Björn Wolfgang Schuller<sup>1,2,3</sup>

<sup>1</sup>CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany

<sup>2</sup>Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

<sup>2</sup>GLAM – Group on Language, Audio, & Music, Imperial College London, UK

## ABSTRACT

The correlation between the sharpness of loss minima and generalisation in the context of deep neural networks has been subject to discussion for a long time. Whilst mostly investigated in the context of selected benchmark data sets in the area of computer vision, we explore this aspect for the acoustic scene classification task of the DCASE2020 challenge data. Our analysis is based on two-dimensional filter-normalised visualisations and a derived sharpness measure. Our exploratory analysis shows that sharper minima tend to show better generalisation than flat minima—even more so for out-of-domain data, recorded from previously unseen devices—, thus adding to the dispute about better generalisation capabilities of flat minima. We further find that, in particular, the choice of optimisers is a main driver of the sharpness of minima and we discuss resulting limitations with respect to comparability. Our code, trained model states and loss landscape visualisations are publicly available.

**Index Terms**— acoustic scene classification, sharp minima, loss landscape, generalisation, deep neural networks

## 1. INTRODUCTION

When training *artificial neural networks* (ANNs) on a specific task, one of the key challenges lies in the network’s ability to generalise to unseen data. As can be interpreted from the universal approximation theorem [1], ANNs are well capable of representing the underlying data distribution of any task. In practice—especially given a network with enough depth—good fits of the training data with converging loss values and perfect evaluation metrics are often easy to find. However, this does not translate to unseen data, as the generalisation error can vary hugely for almost perfect training loss and can be influenced by the amount of training data, the choice of network architecture, optimiser or batch size [2], among other things. Models with a high generalisation gap are considered to be overfitted and often perform even worse if the unseen data is *out-of-domain* (OOD). This can, for instance, be observed in the yearly DCASE *acoustic scene classification* (ASC) challenge, in which the organisers added new recording conditions, such as different recording devices or cities, only to the test data. Critically, model selection, in the form of choosing hyper-parameters or ‘early stopping’, is predominantly performed based on validation performance, which on its own can bring quite some limitations as, for instance, reported for OOD performance [3].

An alternative perspective on model states can be gained by examining the behaviour of loss functions. Specifically, some characteristics of a model state’s minimum have been pointed out to show an important connection to the generalisation error. *Flatness* and

*sharpness* play a particular role here, with flatter minima often believed to have better generalisation [4], at least since the work of Hochreiter and Schmidhuber [5]. Intuitively, these terms are related to the Hessian matrix, which contains all second-order derivatives, at a given point of a function, for all directions and can thus represent the local curvature behaviour of the function. Yet, an undisputed definition of flatness and sharpness in the high-dimensional parameter space of ANNs is still lacking. Nevertheless, several approaches to quantify flatness and sharpness have been developed over the years, but they have failed to paint a complete picture of the generalisation capabilities based on geometry, as a universal correlation between flatness and generalisation has been disputed [6, 7]. In particular authors in [8] claim that the conclusion that flat minima should generalise better than sharp ones cannot be applied as is without further context. Likewise, Andriushchenko et al.[9] recently observed in multiple cases that sharper minima can generalise better in some modern experimental settings.

Arguably, the most impactful sharpness measure, the  $\epsilon$ -sharpness, was introduced by Keskar et al. [2]. It decodes the information from the eigenvalues of the Hessian matrix, while at the same time avoiding the computation-heavy calculation of the Hessian matrix itself. Alternative measures of sharpness include the consideration of local entropy around a minimum [10] or of the size of the connected region around the minimum where the loss is relatively similar [5]. Li et al. [11] however show that a problem in the interpretability of sharpness measures, such as the  $\epsilon$ -sharpness, may lie in the scaling of the weights. An apparent example is optimisers with weight penalties, which enforce smaller parameters, and are thus more prone to disturbance, leading to sharp minima with good generalisation. In order to overcome this limitation, they suggest to use filter-normalisation for the visualisation of loss landscapes and argue that flatter minima in low-dimensional visualisations with filter-normalised directions go hand-in-hand with better generalisation capabilities, even when compared across different ANN architectures. Even though this relationship is made evident in several instances on a qualitative level, a quantitative measure of the sharpness in the context of filter-normalisation and a corresponding analysis are not provided.

Beyond, a core weakness with respect to the universal validity of the results in most previously mentioned contributions is that experiments are limited to established benchmark data sets for image classification, such as CIFAR-10 [12] or ImageNet [13], and should thus be further verified in different research areas and contexts. In this work, we focus on exploring the ASC task of the DCASE2020 challenge, which belongs to the same category of tasks as CIFAR-10 (10-class classification problem), but comprises a different modality (audio instead of images) and more challenges of real-world data.

The DCASE ASC challenge has seen tremendous influence on the computer audition community [14]. The yearly updated data sets have been the basis for ASC studies ranging from the development of new model architectures [15] and the evaluation of model robustness [16, 17], to investigations of fairness in performance amongst different recording devices and locations [3].

In this contribution, we suggest a new approach to quantitatively measure the sharpness of a local minimum—or at least of the neighbourhood of a ‘well-trained’ model state—and find correlations to the generalisability of ASC models. We design our experiments considering different architectures, training parameters, and optimisation algorithms in order to address the following research questions:

- Is the sharpness derived from a two-dimensional filter-normalised visualisation stable across random directions?
- How does the sharpness of ASC models correlate with the generalisation error for *in-domain* (ID) and OOD data?
- Which hyperparameters of model training are drivers for sharp minima?

These investigations might give insights relevant to the selection of models that generalise better to OOD data, as well as drive the understanding of different factors affecting this generalisation for computer audition, which are both important open questions for ASC.

## 2. METHODOLOGY

### 2.1. Filter-Normalisation

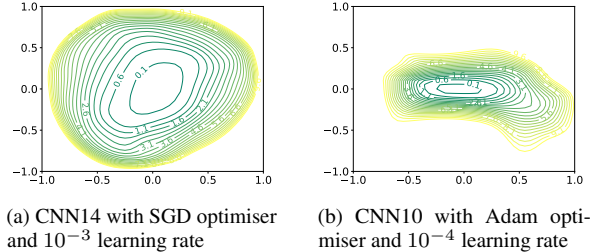
The basis for our characterisation of minima are low-dimensional filter-normalised visualisations of the loss minima as introduced in [11]. The prerequisite for such a visualisation is an ANN with parameters  $\theta$ , which was trained to a model state  $\theta^*$ , close to a local minimum of the loss function, given a training set  $X$ . The precise minimum, however, will most likely not be reached in practice, given a finite time for training, finite numerical precision, and in particular, through techniques such as early stopping. The loss function around the trained model state will nevertheless in most cases increase, when varying any of the parameters  $\theta_i$  of the network. With common ANNs having millions or even billions of parameters, this leads to very high-dimensional loss landscapes. The immediate surroundings of the minimum can best be described with the Hessian matrix. The high dimensionality however makes the calculation of the Hessian matrix very computation-heavy and thus not practical [18], although significant attempts are addressed in this direction [19].

Instead, a common approach to look at the loss landscape is through low-dimensional visualisations. In two dimensions, this can be realised through the choice of random Gaussian vectors  $\delta$  and  $\eta$ , both of the same dimension as  $\theta$ , which are in the following used to project the loss function as

$$f(\alpha, \beta) = L(\theta^* + \alpha\delta + \beta\eta). \quad (1)$$

By varying the scalar variables  $\alpha$  and  $\beta$ , we can depict a 2-dimensional projection of the loss landscape. However, Li et al. point out some weaknesses of the visualisation, as different models—and even different model states of the same architecture—can have differently scaled parameters, thus making them more or less vulnerable to perturbations of the same magnitude [11]. Therefore, they suggest adjusting the perturbations relative to the magnitude of the weights, thus rescaling the random gaussian directions  $\delta$  and  $\eta$  choosing a filter-level normalisation. This can be formulated as

$$\delta_{i,j} \leftarrow \frac{\delta_{i,j}}{\|\delta_{i,j}\|} \|\theta_{i,j}\|, \quad (2)$$



**Fig. 1:** Visualisation of the two-dimensional filter-normalised loss landscape for two different model states with different architectures and training paradigms.

where the indices of  $\delta_{i,j}$  and  $\theta_{i,j}$  refer to the components of  $\delta$  corresponding to the  $j$ th filter of the  $i$ th layer in a convolutional neural network. Figure 1 shows two examples of filter-normalised loss landscapes in 2D around a minimum with  $\alpha$  and  $\beta$  ranging from -1 to 1, thus varying the filters of the network by around  $\pm 100\%$ . We will use plots of this kind for the following analyses with the adapted code provided by the authors in [11]. As the filter-normalised plots are solving the problem of different scales of filters, the authors claim that flatter minima in this representation, despite the heavy reduction in dimensionality, indicate better generalisation, which is underlined with a qualitative analysis of several model states, trained on the CIFAR-10 dataset.

### 2.2. Sharpness

In order to quantitatively evaluate these claims for our ASC problem, we base our analysis on the  $\epsilon$ -sharpness, which is prominently used in the literature. This measure focuses on a small neighbourhood of a minimum and computes the largest value potentially attained by the loss function and is considered a good approximation of the curvature of the minimum and thus, of the sharpness or flatness of the minimum. Formally, it is defined as

$$s_\epsilon = \frac{\max_{\theta \in B(\epsilon, \theta^*)} (L(\theta) - L(\theta^*))}{1 + L(\theta^*)} \times 100, \quad (3)$$

where  $B(\epsilon, \theta^*)$  is a Euclidean ball centred on a minimum  $\theta^*$  with radius  $\epsilon$ , i.e.,  $\{\theta \in \mathbb{R}^n : \|\theta - \theta^*\| < \epsilon\}$ .

We follow (3) to calculate a quantitative sharpness measure of the two-dimensional visualisation (obtained from (1) and (2)). We will utilise this sharpness measure in the following to analyse the influences certain experimental settings have on the sharpness of minima and, further, what sharpness can tell us about the generalisation of an ASC model on unseen data.

## 3. EXPERIMENTS AND DISCUSSION

### 3.1. Dataset

As our dataset, we use the development partition of the DCASE 2020 Acoustic Scene Classification dataset [20] and evaluate the experiments based on the standard metric accuracy, which is defined as the ratio of correctly classified samples over all samples. The dataset includes 64 hours of audio segments from 10 different acoustic scenes, recorded in 10 European cities with 3 real devices (denoted as A, B, C), as well as data from 6 simulated devices (denoted as S1-S6). We use the official training/evaluation splits with devices S4-S6 only appearing in the test set (OOD). The data is evenly distributed across

cities, whereas device A (Soundman OKM II Klassik/studio A3) is dominating over B, C, and the simulated devices. We extract 64-bin log-Mel spectrograms with a hop size of 10 ms and a window size of 32 ms, additionally resampling the 10 s long audio segments to 16 kHz.

### 3.2. Model training

Our initial experiments involved two *convolutional neural network* (CNN)-based architectures, the *pre-trained audio neural networks* (PANNs) CNN10 and CNN14 [21] both with random initialisation and around 5.2 million and 80.8 million parameters, respectively, which have frequently been applied to computer audition tasks, including the DCASE ASC task [3, 21]. Their convolutional nature is well in line with the CNNs for which the filter-normalisation was developed. We explored widely-used optimisers, such as *Adaptive Moment Estimation* (Adam) and *stochastic gradient descent* (SGD) with momentum, as well as less common optimisation algorithms, such as the second-order *Kronecker-factored approximate curvature* (KFAC) [22] and *gradient descent: the ultimate optimiser* (GDTUO) [23]. KFAC utilises approximations to the Hessian matrix to improve convergence speed, while GDTUO automatically adjusts hyperparameters using a stack of multiple optimisers, which in this case involves two stacked Adam optimisers, called hyperoptimisers. However, both KFAC and GDTUO resulted in higher computational costs in terms of runtime and memory requirements per optimisation step. We ran a grid-search for hyperparameters as manifested in Table 1, leading to overall 38 trained model states. Besides the learning rate, we used default parameters for the optimisers, with SGD using a momentum of 0.9.

In all cases, the training was stopped after 50 epochs and the best model state of the epoch with the highest accuracy on the development set used for testing. The training is implemented in PYTORCH 1.13.1+cu117 and models were trained on a NVIDIA GeForce GTX TITAN X and a NVIDIA TITAN X (Pascal), both with 12GB RAM. The training time per epoch mostly varied depending on the chosen optimiser, ranging from approximately four minutes for the SGD and Adam optimisers to slightly over six minutes for KFAC, and up to around 18 minutes for GDTUO. Our code and trained model states are publicly available<sup>1</sup>.

**Table 1:** Overview of the grid search parameters for model training.

Network	CNN10, CNN14
Optimiser	SGD, Adam, GDTUO <sup>2</sup> , KFAC <sup>3</sup>
Learning Rate	$10^{-3}$ , $10^{-4}$ <sup>4</sup> , $10^{-5}$ <sup>5</sup>
Batch Size	16, 32 <sup>6</sup>
Random Seeds	42, 43

### 3.3. On the robustness towards random directions

Even though not emphasised by the authors of the filter-normalisation method, the choice of the random Gaussian direction should have

<sup>1</sup>[https://github.com/EIHW/ASC\\_Sharpness](https://github.com/EIHW/ASC_Sharpness)

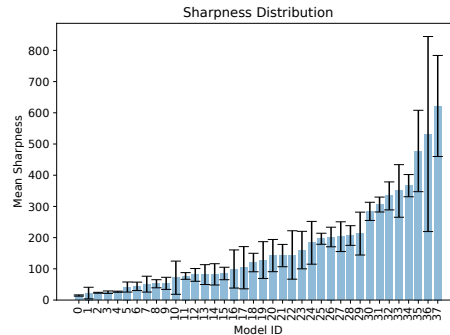
<sup>2</sup>Learning rate refers to the highest optimiser on the stack for GDTUO, since this is not a hyperoptimiser.

<sup>3</sup>GDTUO and KFAC are only applied to the CNN10 architecture, due to hardware limitations.

<sup>4</sup>Not applied to SGD due to suboptimal convergence.

<sup>5</sup>Only applied to KFAC due to suboptimal convergence of other optimisers.

<sup>6</sup>Not applied to KFAC due to hardware limitations.



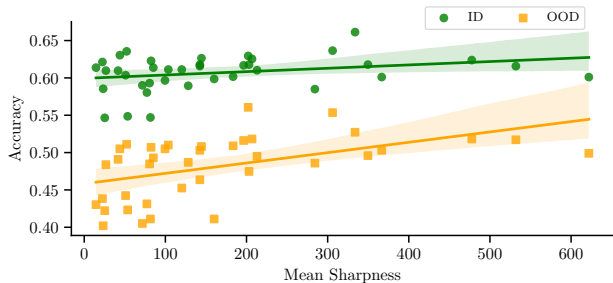
**Fig. 2:** Distribution of sharpness-measures. Each bar indicates the mean sharpness value with the standard deviation of a trained model state in three two-dimensional plots with different random directions.

some impact on the measured or perceived sharpness of a given minimum. To mitigate this impacts in similar settings the authors in [24] use more directions in the parameter space, while in [25], it is suggested to analyse projections along Hessian directions as an alternative method. Nevertheless, most interpretations of the sharpness of minima are limited to (statistics of) a low-dimensional analysis and often show consistent trends across different random directions [26], [27], [28]. We tested the robustness of our sharpness measure by calculating it based on three plots with different random directions. In order to stay in line with the visual argumentation of the plots, as well as the characteristics of the filter-normalisation, we chose a neighbourhood of radius 0.25 to calculate the sharpness. Due to the high computational costs of such visualisations, the resolution was set to 0.025 in each direction, leading to 121 loss values per visualisation. The time required to compute one sharpness value in this scenario is around 45 minutes on a single NVIDIA A40 GPU with 16GB RAM.

Figure 2 shows the mean sharpness and standard deviation for each trained model based on three different plots per model. Most model states show a relatively low standard deviation compared to the mean sharpness, allowing us to further interpret the sharpness in different settings. A few exceptions with high standard deviations indicate some limitations of this approach, which might, however, be mitigated by sampling more sharpness-measures per model. Similar analyses of the stability of sharpness-measures with respect to different random directions have previously been reported [27].

### 3.4. On the impact of sharpness on generalisation

In order to gain insights into the generalisation capabilities of flat and sharp minima in ASC, we plot the test accuracies of the trained model states against their mean sharpness value in Figure 3. We thereby consider the accuracy for ID and OOD separately. To that end, we define OOD performance as the accuracy evaluated on the devices not represented in the training data, namely S4, S5, and S6, whilst ID performance is evaluated on the devices A, B, C, S1, S2 and S3, which are known at training time. Note that all discussed model states show a nearly 100% accuracy on the training data, such that one minus the test accuracy can be interpreted as the generalisation gap. Firstly, we note a tendency that, in our experiments, sharper minima show a better generalisation than flat minima. This is a rather surprising finding, as most of the existing literature reports preferable characteristics of flat minima in the computer vision domain, e.g., [5], [10], [2], [29], [30], [31], [32], whilst only few studies report



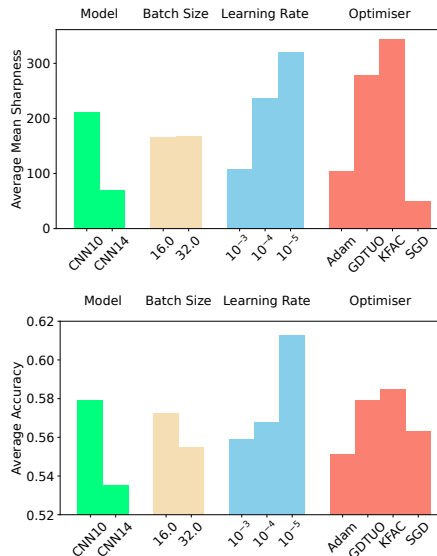
**Fig. 3:** Correlation plot between sharpness of minima (the higher, the sharper) and test accuracy for all trained models. Showing best-fit line and 95% confidence intervals for different models.

on good generalisation in context of sharp minima [33, 9]. Further investigations are necessary to unravel, whether our results are an indication of a general disparity of the impact of sharpness on generalisation in acoustic scene classification and image classification. Critical differences in the learning of computer audition models compared to computer vision models have been reported in our previous work: when fine-tuning a CNN for a computer audition task, the first layers were subject to more changes than the later layers [34]. This finding contradicts the common understanding, resulting from computer vision analyses, of earlier filters being trained to recognise low-complexity objects, such as edges, and are thus transferable without major changes amongst different tasks.

Moreover, this effect seems to be considerably higher for OOD accuracy compared to ID accuracy, as we observe a correlation of .49 in the former and a correlation of .28 in the latter case. Based on our exploratory analysis, we hypothesise that flatter minima are over-optimised for the ID devices –in particular, device A which dominates the training set– and thus fail to generalise well to unseen devices. Nevertheless, the reasons for positive correlations between sharpness and generalisation are not obvious at this moment and should be further looked into.

### 3.5. On the impact of hyperparameters on sharpness

As a final aspect, we analyse the impact of the choice of different hyperparameters or experimental settings on the sharpness and compare these to the corresponding impact on test accuracy. Figure 4 suggests that both sharpness and accuracy are similarly affected by the training parameters. Certain hyperparameters lead to a higher value in both subplots compared to the other hyperparameters in the group, except for the batch size. This result is in line with our previous findings of sharper minima tending to have better generalisation. However, upon closer examination, it becomes apparent that the amount by which both subplots are affected by a certain group can vary considerably, as the selection of optimisers seems to have the highest impact on sharpness, which is not the case for the test accuracy. This provides us with some insights about when a deduction of generalisation from sharpness might be more reasonable, as, for instance, different optimisers seem to bring different tendencies in sharpness, which might not fully translate to generalisation. A remarkable similarity between average mean sharpness and average test accuracy can, however, be observed for the two model architectures, whose sharpness derives from a different(-dimensional) loss landscape. Note that the choice of learning rates and optimisers were not independent of each other, which limits their separate expressiveness.



**Fig. 4:** Disaggregated distribution of mean sharpness and accuracy across hyperparameters. Each bar averages the mean sharpness or accuracy of all trained models states, grouped by the different types of hyperparameters.

### 3.6. Limitations

One of the limitations of our approach lies in the robustness of the sharpness measure, which might, however, be overcome by more efficient implementations, allowing for the consideration of additional random directions. Beyond that, a more thorough analysis of the convergence status of models and its impact on the sharpness measure and generalisation seems desirable. Especially, considering that not all experimental details could be investigated in depth, this contribution can only be a piece in the debate about flat versus sharp minima in ASC in particular and computer audition in general. Beyond, the reasons for good generalisation capabilities of sharp minima in our exploratory study need to be further investigated as the impact of individual hyperparameters on the training needs to be better understood.

## 4. CONCLUSIONS

In this contribution, we explored the sharpness of minima in the loss function for acoustic scene classification models and its impact on the generalisation capabilities in different, practice-relevant, experimental settings. We found that for our trained models, sharper minima generalised better to unseen (in particular to OOD) data, which has rarely been observed in the computer vision domain. Our approach shows some limitations, as for instance, the choice of optimisers has a higher impact on the sharpness of minima than on the generalisation. In future work, we plan to focus on more efficient and interpretable implementations of sharpness measures and to better understand individual effects of hyperparameters before our findings can be put into practice.

## 5. ACKNOWLEDGEMENTS

This work was partially funded by the DFG’s Reinhart Koselleck project No. 442218748 (AUDIONOMOUS).

## 6. REFERENCES

- [1] Kurt Hornik, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [2] 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.
- [3] Andreas Triantafyllopoulos, Manuel Milling, Konstantinos Drossos, and Björn Schuller, "Fairness and underspecification in acoustic scene classification: The case for disaggregated evaluations," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 70–74.
- [4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park, "Swad: Domain generalization by seeking flat minima," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22405–22418, 2021.
- [5] Sepp Hochreiter and Jürgen Schmidhuber, "Flat minima," *Neural computation*, vol. 9, no. 1, pp. 1–42, 1997.
- [6] Shuofeng Zhang, Isaac Reid, Guillermo Valle Pérez, and Ard Louis, "Why flatness does and does not correlate with generalization for deep neural networks," *arXiv preprint arXiv:2103.06219*, 2021.
- [7] Diego Granzio, "Flatness is a false friend," *arXiv preprint arXiv:2006.09091*, 2020.
- [8] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio, "Sharp minima can generalize for deep nets," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1019–1028.
- [9] 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.
- [10] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina, "Entropy-sgd: Biasing gradient descent into wide valleys," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, pp. 124018, 2019.
- [11] 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.
- [12] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [14] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Acoustic scene classification: An overview of dcase 2017 challenge entries," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 411–415.
- [15] Jinxi Guo, Ning Xu, Li-Jia Li, and Abeer Alwan, "Attention based cldnns for short-duration acoustic scene classification," in *Proc. Interspeech 2017*, 2017, pp. 469–473.
- [16] Hu Hu, Chao-Han Huck Yang, Xianjun Xia, Xue Bai, Xin Tang, Yajian Wang, Shutong Niu, Li Chai, Juanjuan Li, Hongning Zhu, et al., "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," *arXiv preprint arXiv:2007.08389*, 2020.
- [17] Lam Pham, Ian McLoughlin, Huy Phan, and Ramaswamy Palaniappan, "A Robust Framework for Acoustic Scene Classification," in *Proc. Interspeech 2019*, 2019, pp. 3634–3638.
- [18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard, "Robustness via curvature regularization, and vice versa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9078–9086.
- [19] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney, "Pyhessian: Neural networks through the lens of the hessian," in *2020 IEEE international conference on big data (Big data)*. IEEE, 2020, pp. 581–590.
- [20] 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.
- [21] 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.
- [22] James Martens and Roger Grosse, "Optimizing neural networks with kronecker-factored approximate curvature," in *International conference on machine learning*, 2015, pp. 2408–2417.
- [23] Kartik Chandra, Audrey Xie, Jonathan Ragan-Kelley, and Erik Meijer, "Gradient descent: The ultimate optimizer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8214–8225, 2022.
- [24] Stefan Horoi, Jessie Huang, Bastian Rieck, Guillaume Lajoie, Guy Wolf, and Smita Krishnaswamy, "Exploring the geometry and topology of neural network loss landscapes," in *Advances in Intelligent Data Analysis XX: 20th International Symposium on Intelligent Data Analysis, IDA 2022, Rennes, France, April 20–22, 2022, Proceedings*. Springer, 2022, pp. 171–184.
- [25] Lucas Bötcher and Gregory Wheeler, "Visualizing high-dimensional loss landscapes with hessian directions," *arXiv preprint arXiv:2208.13219*, 2022.
- [26] Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe, "Qualitatively characterizing neural network optimization problems," *arXiv preprint arXiv:1412.6544*, 2014.
- [27] Dongxian Wu, Shu-Tao Xia, and Yisen Wang, "Adversarial weight perturbation helps robust generalization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2958–2969, 2020.
- [28] Leslie N Smith and Nicholay Topin, "Exploring loss function topology with cyclical learning rates," *arXiv preprint arXiv:1702.04283*, 2017.
- [29] P Izmailov, AG Wilson, D Podoprikin, D Vetrov, and T Garipov, "Averaging weights leads to wider optima and better generalization," in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018, pp. 876–885.
- [30] Haowei He, Gao Huang, and Yang Yuan, "Asymmetric valleys: Beyond sharp and flat local minima," *Advances in neural information processing systems*, vol. 32, 2019.
- [31] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al., "Towards theoretically understanding why sgd generalizes better than adam in deep learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21285–21296, 2020.
- [32] David Stutz, Matthias Hein, and Bernt Schiele, "Relating adversarially robust generalization to flat minima," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7807–7817.
- [33] Enzo Tartaglione, Andrea Bragagnolo, and Marco Grangetto, "Pruning artificial neural networks: A way to find well-generalizing, high-entropy sharp minima," in *Artificial Neural Networks and Machine Learning—ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part II 29*. Springer, 2020, pp. 67–78.
- [34] Andreas Triantafyllopoulos and Björn W Schuller, "The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7268–7272.