TOWARDS BRAIN INSPIRED DESIGN FOR ADDRESSING THE SHORTCOMINGS OF ANNS

Fahad Sarfraz[1,2], Bahram Zonooz[1,*]& Elahe Arani[1,3,*]
[1] Dep. of Mathematics and Computer Science, Eindhoven University of Technology (TU/e), Netherlands
[2] TomTom, Netherlands
[3] Wayve Technologies Ltd, London, United Kingdom
{f.sarfraz, b.zonooz, e.arani}@tue.nl

ABSTRACT

As our understanding of the mechanisms of brain function is enhanced, the value of insights gained from neuroscience to the development of AI algorithms deserves further consideration. Here, we draw parallels with an existing tree-based ANN architecture and a recent neuroscience study (Shadmehr, 2020) arguing that the error-based organization of neurons in the cerebellum that share a preference for a personalized view of the entire error space, may account for several desirable features of behavior and learning. We then analyze the learning behavior and characteristics of the model under varying scenarios to gauge the potential benefits of a similar mechanism in ANN. Our empirical results suggest that having separate populations of neurons with personalized error views can enable efficient learning under class imbalance and limited data, and reduce the susceptibility to unintended shortcut strategies, leading to improved generalization. This work highlights the potential of translating the learning machinery of the brain into the design of a new generation of ANNs and provides further credence to the argument that biologically inspired AI may hold the key to overcoming the shortcomings of ANNs.

1 INTRODUCTION

Artificial neural networks (ANNs) have achieved remarkable performance on many vision tasks which have been enabled by the considerable progress in developing deeper and more complex network architectures. However, despite the performance gains, the existing networks have been shown to be brittle and have several limitations and shortcomings. They require huge amounts of data to train, struggle with noisy and imbalanced datasets, do not generalize well to out-of-distribution data, and are vulnerable to shortcut learning and adversarial attacks (Vandervert, 2016). While there have been studies on addressing these challenges individually, the majority of these specialized techniques and regularization approaches for overcoming a specific challenge lead to a trade-off in performance and do not provide a general solution (Tsipras et al., 2019).

Humans, on the other hand, excel at learning efficiently even under challenging scenarios with limited data and can generalize well to novel scenarios. Neuroscience has made substantial progress in understanding the mechanisms of brain functions and the design principles it employs to enable efficient learning (Hassabis et al., 2017; Kudithipudi et al., 2022; Hawkins, 2021; Macpherson et al., 2021). It is therefore, important to further exploit insights from our enhanced understanding of the learning machinery of the brain into the development of AI algorithms.

We consider a recent study by Shadmehr (2020) that examines the organization of neurons in the cerebellum, an important learning site in the brain, and resembles a three-layer feedforward network (see Figure 1). The neurons in the middle layer of the cerebellum are grouped into small populations that receive a personalized view of the entire error space. This is in stark contrast to standard ANNs which lack any such organization of neurons and each unit in the network receives the same error signal. Therefore, we attempt to study the potential effect of a similar error-dependent organization of neurons in ANNs. To this end, in the object recognition task in ANNs, we consider the classification error associated with a learned semantic grouping of object classes as partial views of the error

[*]Equal advisory role.

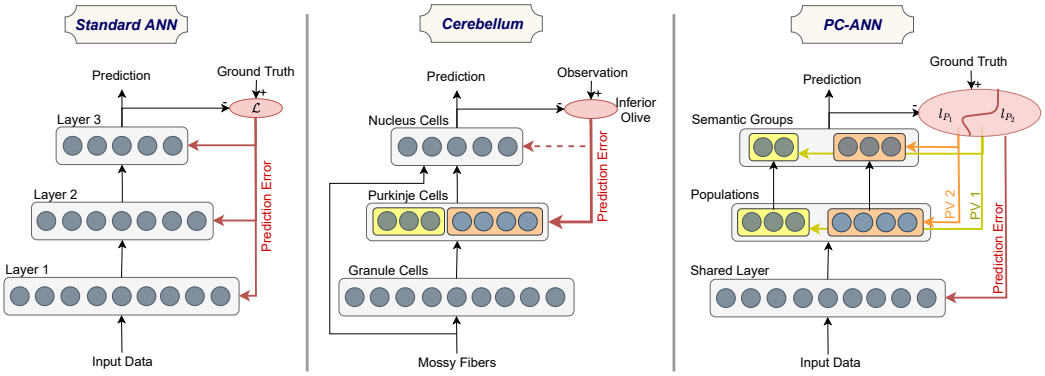

Figure 1: In *standard ANN*, there is no grouping of neurons within a layer, and all neurons receive the same global error signal for weight updates. In contrast, The *Cerebellum* has a vastly different design and learning mechanism. It resembles a three-layer feedforward network where the inferior olive organizes Purkinje cells in the hidden layer into small populations which receive a limited personalized view of the error space. Similar to the cerebellum, the desired population coding based ANN (*PC-ANN*) would form an error-dependent grouping of neurons into populations that learn from partial error views from the classification error associated with learned semantic groupings of classes. We consider SplitNet as an instance of an architecture that bears similarities PC-ANN.

space and the corresponding set of disjoint subnetworks as populations that share a preference for a particular partial error view. From this perspective, our intended learning paradigm can be more aligned with tree-structured ANNs.

Thus, we consider SplitNet (Kim et al., 2017), originally proposed to improve inference speed and reduce the number of parameters, as a suitable tree-structured method to assess the potential of population coding in ANNs as we see many similarities between its network design and the error-dependent organization of neurons in the cerebellum (see Figure 1). Analogous to the grouping of neurons in the cerebellum, SplitNet learns to split the network weights into multiple groups that use disjoint sets of features. In particular, since the logit values associated with the semantically disparate class groups only depend on the corresponding branched subtrees of the network and not the other subtrees, each group (subtrees) receives a gradient signal which is biased towards correcting the error associated with their corresponding semantic group (partial error view), similar to how populations in the cerebellum share preference for a biased error view. Finally, similar to the cerebellum, which receives a highly processed input, SplitNet has a shared layer that extracts features from the input data before splitting them into separate populations. Therefore, we consider SplitNet to bear some similarities to the population coding in the cerebellum and is, therefore, suitable for conducting our empirical study.

We assess the potential benefits of the error-based organization of neurons in the design of ANNs under varying training conditions and assess its effect on the learned model. Our empirical evaluation demonstrates the effectiveness of the considered architecture in improving the generalization of the model over standard training under challenging scenarios. It provides considerable performance gains under class imbalance which is inherent in real-world datasets and significantly improves the sample efficiency of the model, enabling the model to generalize better with fewer data. Additionally, our empirical results suggest that error-based organization of neurons can reduce the texture bias and vulnerability to unintended shortcut cues which improve generalization to out-of-distribution data. We attribute these improvements to the flexibility of the subnets to explore the feature space more and learn specialized features for the semantic groups. Furthermore, our analyses of the characteristics of the model suggest that it compresses more information and converges to flatter minima. Notably, all of these benefits come merely from the design of the network rather than explicit regularization or specific techniques for each scenario. Our empirical results highlight the potential of error-based grouping and partial error views based learning mechanisms in ANNs.

Our work aims to bring the perspective of population coding-based design in ANNs and presents it as a promising direction for further research. We believe that exploring the design space of population

coding-based ANNs can lead to more reliable and robust models that may address some of the key challenges and limitations of current AI models.

2 BACKGROUND AND METHODOLOGY

We outline the background and premises of our study: Section 2.1 reviews population coding in the cerebellum (Shadmehr, 2020), Section 2.2 details SplitNet, and Section 2.3 explores its parallels with ANNs and the benefits of population coding.

2.1 POPULATION CODING IN THE CEREBELLUM

The cerebellum is an important learning site in the brain (Rapoport et al., 2000; Vandervert, 2016) and, therefore, several studies in neuroscience have scrutinized how efficient learning is enabled in the cerebellum (Herzfeld et al., 2020; 2018; Kitazawa et al., 1998; Kojima et al., 2010). It has a relatively simple circuit architecture that resembles a three-layer feedforward network of neurons in which the "hidden layer" consists of Purkinje cells (P-cells), and the output layer consists of deep cerebellar nucleus (DCN) neurons. Our study focuses on the recent work of Shadmehr (2020), which provides an extensive overview of the learning characteristics and organization of neurons in the cerebellum from a machine learning perspective and its implications. Unlike an ANN, P-cells are grouped into small populations that converge on single DCN neurons. Furthermore, the error signal conveyed to the P-Cells, which in turn act as surrogate teachers for the downstream DCN neurons they project to, is not a fair reflection of the entire error space, but is rather biased to provide a limited (personalized) view of the error space. This error-dependent grouping of P-cells into populations is believed to play a crucial role in enabling efficient learning in the cerebellum. Our study aims to bring this perspective to ANNs and to study the potential benefits of such an architecture.

2.2 SPLITNET

To this end, we consider an existing ANN architecture that bears some resemblance to a grouping of neurons with personalized error views. SplitNet (Kim et al., 2017) was originally proposed to optimize the inference speed of the model by learning a tree-structured network architecture that is highly parallelizable. The method involves splitting the network into a set of subnetworks that share a common lower layer and using a disjoint set of features for the specific group of classes associated with the subnetwork. SplitNet employs a two-stage learning scheme whereby in the first stage classes-to-group and features-to-group assignment matrices are learned along with the network weights while regularizing them to be disjoint across groups. The learned assignment matrices are then utilized to obtain a tree-structured network that involves no connection between branched subtrees of semantically disparate class groups, which are subsequently finetuned with the cross-entropy loss.

Concretely, for a given number of groups, G, the vector of assignment of the feature group and the vector of assignment of the class group for the group g ($1 \leq g \leq G$) are given by $p_g \in R^D$ and $q_g \in R^K$ where D is the dimension of the features and K is the number of classes. p_g and q_g define a group together, where p_g represents the features associated with the group and q_g indicates a set of classes assigned to the group. The disjoint set of classes and features are learning by imposing a constraint on the network weight at each layer W^l to be a block-diagonal matrix, where each block W_g^l is associated with a class group $g \in G$. There is no overlap between groups, either in features or classes, so each disjoint group of classes has exclusive features associated with it. See Appendix for regularization objectives.

2.3 STUDYING THE POTENTIAL BENEFITS OF POPULATION CODING IN ANNS

The resemblance of the cerebellum to a feedforward network and a preliminary understanding of the error-driven organization of neurons and the learning mechanisms it employs provide us with an opportunity to study the benefits of such an architecture in ANNs. Standard learning consists of evaluating an overall error term (e.g. mean cross-entropy loss over a training batch) and subsequently updating each neuron's weight in the gradient direction, which minimizes the loss term. As explained in Section 2.1, this is in stark contrast to how the cerebellum learns, and therefore we aim

Table 1: Test accuracy on different datasets. PC-ANN consistently improves the generalization of the model across datasets of varying complexity, demonstrating its versatility.

	CIFAR-10	CIFAR-100	TINY-IMAGENET
BASELINE	92.49 $\pm$0.25	73.65 $\pm$0.18	49.14 $\pm$0.49
PC-ANN	**93.24** $\pm$0.21	**75.33** $\pm$0.47	**53.02** $\pm$0.22

to study the potential impact of a similar error-dependent grouping of neurons into populations and subsequently learning from partial error views in ANNs.

To this end, we first define the framework within which we conduct our study by drawing parallels under the classification task in ANNs. We aim to learn semantically disparate grouping of object classes which can be represented by a disjoint set of features. Semantically similar classes are likely to share features and meaningfully partition the input space. Therefore, the classification error associated with each semantic group can provide a personalized view of the error space, which can be subsequently utilized to learn specialized features in the associated subnetwork (population of neurons). This paradigm naturally lends itself to tree-structured network architectures such as SplitNet. Figure 1 shows the similarities between the cerebellum and structure and the learning dynamics of SplitNet (referred to as PC-ANN for emphasis). Notably, a closer look at the backpropagation of errors in SplitNet reveals an intriguing property that makes it suitable for our study as an instance of ANN architecture that bears similarity to the population coding in the cerebellum: the logits for each class in a semantic group depend only on the associated subtree (population) which therefore receives an error signal which is biased towards correcting the error associated with the semantic group (partial error). For instance, consider the scenario where we have two semantic groups: living and non-living, and the input image is of a cat. The logit values for non-living classes are provided by the associated subtree and vice versa. Hence, as the error signal for each unit depends on its involvement in the forward pass, the subtree for the living semantic group will receive an error signal biased towards correcting the error associated with the logits for living classes and similarly for the non-living subtree. Therefore, we posit that splitNet implicitly utilizes partial error views to create specialized populations of neurons. Studying the performance and characteristics of such a network enables us to gauge the potential benefits of mimicking population coding in ANNs.

3 EMPIRICAL EVALUATION

To study the potential benefits of incorporating a similar mechanism for population coding in ANN, we evaluate the characteristics and learning behavior of SplitNet in various challenging scenarios. Therefore, we refer to SplitNet as an instance of the desired population coding-based ANN (PC-ANN), the subnetworks as populations, and the classification error associated with the learned semantic grouping of classes as partial error views to emphasize our focus on studying the potential impact of a similar mechanism of population coding in ANNs as the cerebellum.

3.1 PERFORMANCE

To assess model versatility, we evaluate multiple datasets of varying complexity. Table 1 shows that PC-ANN consistently improves generalization, particularly on complex datasets with more classes and higher interclass similarity. These results suggest that PC-ANN efficiently forms semantic groups and learns from partial error views. By providing targeted signals to specific populations, partial error views encourage broader feature exploration and specialized learning, helping mitigate narrow learning pitfalls (Tramèr et al., 2020).

3.2 OUT-OF-DISTRIBUTION (OOD) GENERALIZATION

A long-standing challenge for AI is its inability to generalize well to OOD data, while humans excel at generalizing to novel situations. To test whether population coding enables the model to learn more generalizable features, we consider two challenging scenarios. We first utilize the cleaned version of the DomainNet dataset (Peng et al., 2019) that consists of data from different domains on 345 object classes. We train the models on the real domain and use the painting, clip art, sketch,

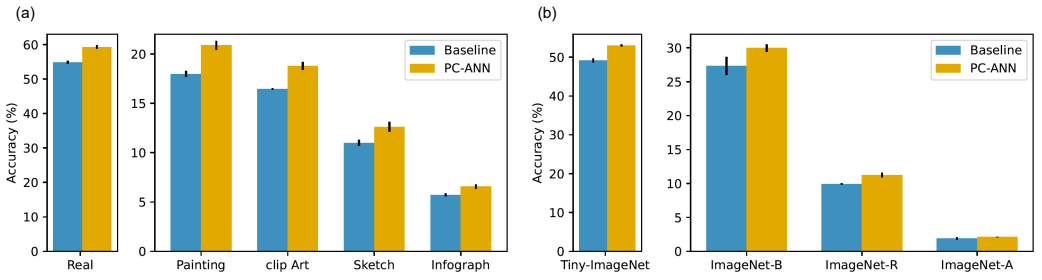

Figure 2: Comparison of OOD Generalization: (a) performance of models trained on real photos and tested on various out-of-distribution domains. (b) Generalization of models trained on Tiny-ImageNet and tested on common classes in different variants of ImageNet dataset. Population coding enables the model to learn better generalizable features, improving OOD generalization.

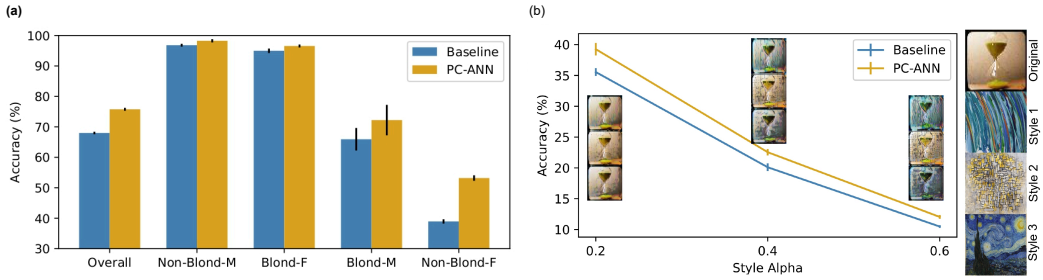

Figure 3: (a) Shortcut learning analysis on CelebA-Skewed: PC-ANN significantly reduces reliance on the blonde color shortcut, as reflected in the performance gap between Blonde-M and Non-Blonde-F. (b) Texture bias analysis on Tiny-ImageNet: PC-ANN achieves better generalization on stylized images, indicating lower texture bias than standard ANNs.

and infograph domains for our OOD testing. We also consider different variants of the ImageNet dataset (Deng et al., 2009). ImageNet-R (Hendrycks et al., 2021) and ImageNet-B (Hendrycks et al., 2021) contain images from 11 different renditions and real blurry images from a subset of 100 classes of ImageNet classes, respectively. However, ImageNet-A (Hendrycks et al., 2019b) provides a dataset for naturally occurring adversarial examples. We test the models trained on Tiny-ImageNet datasets on the common subset of classes within each of these ImageNet variants for OOD evaluation. PC-ANN provides better generalization across all the domains of DomainNet (Figure 2 (a). We observe similar gains on different variants of the ImageNet datasets (Figure 2 (b)). Although the difference in ImageNet-A is minor, it provides early evidence that having separate subnetworks may improve adversarial robustness. We hypothesize that the generalization gains with PC-ANN may be attributed to learning a specialized set of features for the learned semantic groups.

3.3 Shortcut Learning

Shortcuts are decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, including real-world scenarios (Geirhos et al., 2020). As the models are typically trained to maximize the training accuracy, they are quite likely to rely on spurious correlations: associations that are predictive of the training data, but not valid for the actual task. A major challenge for enabling efficient learning in ANNs is therefore to control the sensitivity of the training process to such spurious correlations. To evaluate the susceptibility of the model to shortcut learning, we follow the analysis in (Jain et al., 2021) and consider a gender classification task based on CelebA dataset (Liu et al., 2015) (CelebA-Skewed) where the training dataset is biased so that it only contains blonde females and non-blond males. Therefore, hair color is highly predictive on training data but not in test data where hair color and gender are independent attributes. Therefore, this may result in a decision rule based only on hair color.

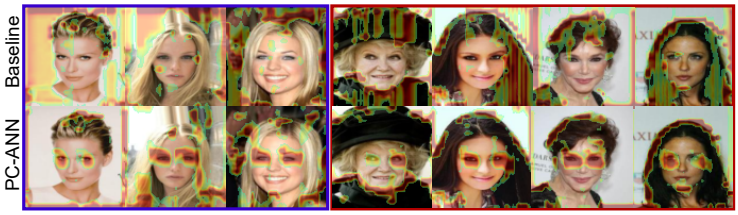

Figure 4: Visual explanations of the models trained on CelebA-Skewed. Attention maps suggest that PC-ANN relies on the salient features of the face to predict gender, while standard training relies more on the unintended shortcut cue (blond hair color in this case).

Figure 3 shows that PC-ANN is in fact less vulnerable to shortcut learning and significantly improves generalization compared to standard ANN. Particularly, we see considerable gain in generalization to non-blond females and blond males without any explicit regularization. To better understand how the two models make decisions, we use the Grad-CAM (Selvaraju et al., 2017) approach to examine the visual explanations of the models. We use the penultimate layer to extract the feature embeddings and use a threshold of 0.4 on the attention maps. Figure 4 shows that population coding remarkably enables the model to attend to the salient features of the face to distinguish between the genders, while standard training relies more on the unintended shortcut cue of the hair color, thus paying more attention to the hair region to inform its decision.

3.4 Texture Bias

Geirhos et al. (2018) conducted a comparative study on convolutional neural networks (CNNs) and human observers on images with a texture-shape cue conflict. Their study revealed that, in sharp contrast to humans, CNNs are strongly biased towards recognizing texture instead of shape, alluding to fundamentally different classification strategies. They further showed that models that learn shape-related features are more robust and generalizable whereas models that rely on texture are susceptible to shortcut learning and result in poor generalization. As PC-ANNs bear some similarities to population coding in the cerebellum, we aim to investigate whether they exhibit behavior that is closer to humans than standard ANNs.

Following Geirhos et al. (2018), we evaluate the texture bias of the model by applying style transfer (Huang & Belongie, 2017) to the Tiny-ImageNet test images. We use four different style images and apply style transfer with varying strengths, i.e. alpha $\in [0.2, 0.4, 0.6]$ so that only the shape of the image corresponds to the correct label. Figure 3 (b) shows that PC-ANN generalizes better under varying stylization strengths, suggesting that it is less biased toward the texture of the image.

Additionally, Section B shows that PC-ANN considerably increases the robustness of model to class imbalance, enhances the sample efficiency and leads to convergence to flatter minima and compresses more information than standard ANN. These results further support our hypothesis that incorporating the details of the learning machinery of the brain can address some of the shortcomings of standard ANNs and bring us closer to human intelligence.

4 Conclusion

We conducted an empirical study to explore the potential benefits of drawing insights from neuroscience findings to the development of AI algorithms. Here, we focused on the recent study (Shadmehr, 2020), which explains the error-based organization of neurons in the cerebellum from a machine learning perspective and attempted to draw parallels with an existing tree-structured ANN. Our empirical evaluation of the considered architecture shows improved robustness to class imbalance and shortcut learning, efficient learning under limited data, and reduced texture bias. Furthermore, the characteristic analyses demonstrate that it compresses higher information in the hidden states, and converges to flatter minima. We hypothesize that these benefits are a consequence of the architecture that resembles population coding in the cerebellum, and further work to explicitly mimic the error-based grouping of neurons in ANN is a promising research direction.

REFERENCES

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*, 2019(12): 124018, 2019.

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*, pp. 248–255. Ieee, 2009.

Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1367–1381, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Jeff Hawkins. *A thousand brains: A new theory of intelligence*. Hachette UK, 2021.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, 2019a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples.(2019). *arXiv preprint cs.LG/1907.07174*, 2019b.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

David J Herzfeld, Yoshiko Kojima, Robijanto Soetedjo, and Reza Shadmehr. Encoding of error and learning to correct that error by the purkinje cells of the cerebellum. *Nature neuroscience*, 21(5): 736–743, 2018.

David J Herzfeld, Nathan J Hall, Marios Tringides, and Stephen G Lisberger. Principles of operation of a cerebellar learning circuit. *Elife*, 9:e55217, 2020.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Saachi Jain, Dimitris Tsipras, and Aleksander Madry. Combining diverse feature priors. *arXiv preprint arXiv:2110.08220*, 2021.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Juyong Kim, Yookoon Park, Gunhee Kim, and Sung Ju Hwang. Splitnet: Learning to semantically split deep networks for parameter reduction and model parallelization. In *International Conference on Machine Learning*, pp. 1866–1874. PMLR, 2017.

Shigeru Kitazawa, Tatsuya Kimura, and Ping-Bo Yin. Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature*, 392(6675):494–497, 1998.

Yoshiko Kojima, Robijanto Soetedjo, and Albert F Fuchs. Changes in simple spike activity of some purkinje cells in the oculomotor vermis during saccade adaptation are appropriate to participate in motor learning. *Journal of Neuroscience*, 30(10):3715–3727, 2010.

Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022.

Alex Lamb, Vikas Verma, Juho Kannala, and Yoshua Bengio. Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 95–103. ACM, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Tom Macpherson, Anne Churchland, Terry Sejnowski, James DiCarlo, Yukiyasu Kamitani, Hidehiko Takahashi, and Takatoshi Hikida. Natural and artificial intelligence: A brief introduction to the interplay between ai and neuroscience research. *Neural Networks*, 144:603–613, 2021.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Mark Rapoport, Robert van Reekum, and Helen Mayberg. The role of the cerebellum in cognition and behavior: a selective review. *The Journal of neuropsychiatry and clinical neurosciences*, 12 (2):193–198, 2000.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. IEEE, 2017.

Reza Shadmehr. Population coding in the cerebellum: a machine learning perspective. *Journal of Neurophysiology*, 124(6):2022–2051, 2020.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2015.

Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning*, pp. 9561–9571. PMLR, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, number 2019, 2019.

Larry Vandervert. The prominent role of the cerebellum in the learning, origin and advancement of culture. *Cerebellum & Ataxias*, 3(1):1–13, 2016.

Fupin Yao. Machine learning with limited data. *arXiv preprint arXiv:2101.11461*, 2021.

Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018.

A REGULARIZATION OBJECTIVES IN SPLITNET

The regularization which assigns features and classes into disjoint groups consists of three objectives:

- Group Weight Regularization, R_W prunes out inter-group connections to obtain block-diagonal weight matrices by minimizing the off-block-diagonal entries;

$$R_W(W,P,Q) = \sum_g \sum_i \|((\mathbb{I}-P_g)WQ_g)_{i*}\|_2 + \sum_g \sum_j \|(P_g W(\mathbb{I}-Q_g))_{*j}\|_2 \tag{1}$$

where $P_g = diag(p_g)$ and $Q_g = diag(q_g)$ are the feature and class group assignment matrices for group g, and $(M)_{i*}$ and $(M)_{*j}$ denote i-th row and j-th column of M. Eq. 1 imposes row/column-wise $l_{2,1}$-norm on the inter-group connections.

- Disjoint Group Assignment, R_D ensures that the group assignment vectors are mutually exclusive by enforcing orthogonality;

$$R_D(P,Q) = \sum_{i<j} p_i \cdot p_j + \sum_{i<j} q_i \cdot q_j \tag{2}$$

- Balanced Group Assignment, R_E encourages the group assignments to be uniformly distributed by minimizing the squared sum of elements in each group assignment vector.

$$R_E(P,Q) = \sum_g \Big((\sum_i p_{gi})^2 + (\sum_j q_{gj})^2 \Big) \tag{3}$$

Therefore, the overall regularization loss is as follows;

$$\Omega(W,P,Q) = \lambda_1 R_W(W,P,Q) + \lambda_2 R_D(P,Q) + \lambda_3 R_E(P,Q) \tag{4}$$

where λ_1, λ_2, and λ_3 control the strength of each regularization. For more details, see Kim et al. (2017).

B ADDITIONAL EMPIRICAL EVALUATION

B.1 IMBALANCED DATASETS

The majority of the benchmark datasets have a uniform distribution of samples across the object classes. However, class imbalance is naturally inherent in the real world, where some objects are more prevalent than others, or it is relatively easier to obtain more data for certain classes. Standard training exhibits bias towards the prevalent classes at the expense of minority class (Dong et al., 2018) leading to a significant drop in generalization performance. While several approaches have been proposed for efficiently training models under class imbalance (Johnson & Khoshgoftaar, 2019) which employs specialized techniques for tackling class imbalance or making certain assumptions about the distribution of data, we still lack a general method that improves the robustness of the underlying learning paradigm.

To evaluate the robustness of PC-ANN to class imbalance, we simulate varying degrees of class imbalance on different datasets. We follow Hendrycks et al. (2019a) and employ the power law model in which the number of training samples for a class c is given by $n_c = \lfloor a/((c-1)^{-\gamma} + b) \rceil$, where $\lfloor . \rceil$ is the integer rounding function, γ represents an imbalance ratio, a and b are offset parameters to specify the largest and smallest class sizes. The training data becomes a power-law class distribution as the imbalance ratio γ decreases. We compare the performance of PC-ANN with the standard ANN on varying degrees of class imbalance $\gamma \in \{2.0, 1.0, 0.6, 0.20\}$ as the γ value decreases, the class imbalance increases. (a, b) are set so that the maximum and minimum class counts are (5000, 250) for CIFAR-10, (500, 25) for CIFAR-100, and Tiny-ImageNet.

Table 2 shows that PC-ANN consistently provides a considerable performance improvement over standard ANNs, especially for more complex datasets with high degrees of class imbalance without any explicit regularization. We believe that a major shortcoming of standard ANNs is that there is no division or specialization of neurons as each unit is involved in correcting the prediction for

Table 2: Comparison of models trained under various levels of class imbalance. Note that the degree of imbalance increases as γ reduces. PC-ANN provides consistent generalization gains over baseline under varying degrees of class imbalance, particularly for higher imbalance on more complex datasets.

	CIFAR-10		CIFAR-100		TINY-IMAGENET	
γ	BASELINE	PC-ANN	BASELINE	PC-ANN	BASELINE	PC-ANN
2	78.02 ±0.68	**79.85** ±0.61	47.42 ±0.49	**51.55** ±0.29	23.67 ±0.08	**28.46** ± 0.07
1	74.59 ±0.42	**75.84** ±0.85	36.87 ±0.34	**42.35** ±0.45	17.74 ±0.29	**21.67** ±0.17
0.6	72.84 ±0.68	**74.66** ±0.51	34.02 ±0.46	**38.41** ±0.43	15.75 ±0.79	**19.54** ±0.16
0.2	71.68 ±0.24	**73.57** ±0.90	30.43 ±0.81	**35.10** ±0.49	13.57 ±0.42	**17.45** ±0.12

Table 3: Performance of the models trained on a different percentage of the training samples. PC-ANN improves the sample efficiency of the model, allowing it to achieve higher performance with less amount of training data.

SAMPLES (%)	CIFAR-10		CIFAR-100		TINY-IMAGENET	
	BASELINE	PC-ANN	BASELINE	PC-ANN	BASELINE	PC-ANN
100	92.49 ±0.25	**93.24** ±0.21	73.67 ±0.18	**75.33** ±0.47	49.14 ±0.49	**53.02** ±0.22
50	88.69 ± 0.23	**90.34** ±0.16	65.22 ±0.21	**68.35** ±0.37	40.27 ±0.30	**46.07** ±0.12
20	80.89 ±0.24	**84.02** ±0.35	48.93 ±0.55	**53.42** ±0.54	26.04 ±1.00	**32.73** ±0.46
10	73.13 ±0.42	**76.36** ±0.43	35.58 ±0.39	**40.95** ±0.34	18.66 ±0.37	**23.09** ±0.13
5	63.48 ±0.86	**67.28** ±0.09	23.77 ±0.42	**28.07** ±0.65	11.83 ±0.45	**15.21** ±0.23
1	42.64 ±0.20	**44.03** ±0.16	8.52 ±0.27	**9.47** ±0.10	4.35 ±0.23	**4.83** ±0.05

every input. Therefore, an imbalanced batch significantly affects the performance of the model as the entire network is adjusted to reduce the loss on the imbalanced batch, thus preferring the dominant class at the expense of less sampled classes. On the contrary, the partial error views and disjoint subtrees in PC-ANN provide more protection to parts of the network, providing implicit regularization. Furthermore, it can take the prevalence of classes into account while grouping them to mitigate the impact of dominant classes on the performance of minority classes, which builds robustness into the learning framework itself.

B.2 SAMPLE EFFICIENCY

Learning complex concepts with a few examples is a hallmark of human intelligence (Zador, 2019), whereas it remains a challenge for ANNs that are data-hungry and require an abundant amount of labeled data to generalize well (Deng et al., 2009). This limits their application in a limited data regime (Yao, 2021). We believe that mimicking the learning machinery of the brain may lead to models that can generalize better under a low data regime. To this end, we compare the performance of the models trained on a subset of different datasets where we only use $p \in [1, 5, 10, 20, 50]$ percentage of the training dataset and test on the full test set. Table 3 shows that PC-ANN consistently provides better generalization compared to standard ANNs, suggesting that it can learn efficiently with limited data. Notably, the performance gains are higher for complex datasets, where both the number of classes and their interclass similarities are higher. We hypothesize that the grouping of neurons into populations allows each population to explore different regions in the feature space, enabling the model to learn more efficiently from partial error views of fewer data.

B.3 CONVERGENCE TO FLATTER MINIMA

As the loss landscape of DNN's optimization objective is non-convex, there can be multiple solutions that can fit the training data, some solutions, however, generalize better because of being in wider valleys where the model predictions do not change drastically with small perturbations in the parameter space compared to the narrow crevices (Selvaraju et al., 2017; Chaudhari et al., 2019; Keskar et al., 2017). To assess whether PC-ANN converges to wider minima, we follow the analysis in Zhang et al. (2018) and add independent Gaussian noise of increasing strength to the parameters

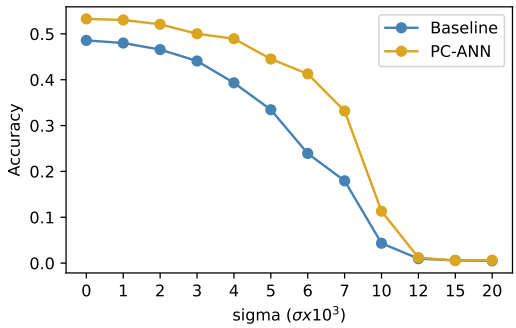

Figure 5: Training accuracy of models on Tiny-ImageNet under varying degrees of weight perturbations. PC-ANN is more stable to weight perturbations, indicating convergence to flatter minima.

Table 4: Comparative analysis on the degree to which a model with frozen learned representations can fit random binary labels. Lower training accuracy indicates higher information compression.

	CIFAR-10	CIFAR-100	TINY-IMAGENET
BASELINE	51.80 ±1.08	92.50 ±2.57	74.37 ±3.89
PC-ANN	**51.60** ±0.65	**90.58** ±2.21	**71.33** ±0.78

of the trained model and analyze the generalization of the trained models on the training dataset. Figure 5 shows that the performance is more stable to the perturbations, suggesting convergence to wider minima.

B.4 INFORMATION COMPRESSION

A number of studies that view ANNs from an information theory perspective (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017) relate the degree to which ANNs compress the information in their hidden states to the bounds on generalization, with higher information compression leading to a stronger bound. To evaluate the effect of population coding on the compression of information in the learned representation, we follow the analysis in Lamb et al. (2019) by freezing the learned representation of the model and measuring how well the frozen representations can fit random labels. we add a 2-layer multi-layer perceptron (MLP) network with 400 and 200 neurons on top of the frozen models trained on the different datasets and fit them on random binary labels. Table 4 shows that PC-ANN enables higher information compression suggesting that the disjoint set of features in PC-ANN allows the model to learn optimal representations that can compress higher semantic information.

C EXPERIMENTAL SETUP

To ensure a fair comparison, we compare the standard training and population coding based training paradigm under uniform experimental settings. Following Kim et al. (2017), we employ WRN-16-8 (Zagoruyko & Komodakis, 2016) for both baseline (Standard-ANN) and SplitNet experiments. Unless otherwise stated, we use the following learning scheme: random horizontal flip and random crop data augmentations with reflective padding of 4 and mean standard normalization; Adam optimizer with $5e^{-4}$ weight decay; 100 epochs; the batch size of 64; and an initial learning rate of $1e^{-4}$, decayed by a factor of 0.1 at epochs 10, 30 and 50. For SplitNet, we use a 2-way split (i.e. $G = 2$) at the final linear layer. For all our experiments, we use $\lambda_1 = 1$, $\lambda_2 = 2$ and $\lambda_3 = 10$. For evaluation, we report the mean and one standard deviation of 3 runs with different seeds.

D LIMITATIONS AND FUTURE WORK

Our study focused on the object recognition task where a meaningful semantic grouping of the classes is possible and utilizes an existing suitable tree-based architecture. As such, the network does not explicitly mimic population coding in the cerebellum, and it is not straightforward to employ it for other tasks (e.g. regression) or when semantic grouping is not possible. We hope that our study inspires exploration of this idea in different domains.

Some potential focus areas for future work can be better strategies for forming error-based groupings of neurons and partial error views and aligning them to minimize the global task error, intertwining the population formation and learning from partial views instead of two separate stages of learning, and explicitly biasing the update rule of the population towards the partial views while also varying the strength of the update in different layers.