ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Towards Understanding The Winner-Take-Most Behavior Of Neural Network Representations

Gilles Peiffer* and Simon Carbonnelle and Christophe De Vleeschouwer

Université catholique de Louvain - ICTEAM, Louvain-la-Neuve, Belgium

Abstract. We analyze neuron-level representations of generalizing and memorizing networks, using a synthetic dataset designed through aggregating hidden patterns into supervision classes. We observe that the average pre-activation of the most activated patterns of a class (and inversely) in each neuron increases during training: a winner-take-most phenomenon. The network applies a divide-and-conquer strategy, where each neuron specializes in classifying different patterns of a class. Through an ablation study, describe three necessary conditions for this phenomenon. Finally, we provide intuition for why it occurs, drawing links with existing work on sample difficulty, gradient coherence, and implicit clustering.

1 Introduction

Despite SGD being a global optimization algorithm, hidden neurons exhibit localized, interpretable behavior [1, 2, 3, 4]. We study this behavior by analyzing hidden neurons on synthetic datasets with simple patterns. Our experiments reveal a winner-take-most (WTM) phenomenon: training amplifies pre-activations for the most activated clusters of a class while suppressing others, often differentiating clusters within the same class strongly. This suggests a divide-and-conquer strategy where neurons specialize in different clusters, linking noise robustness to generalization and feature emergence in vision models.

We work in a simplified setup, allowing us to interpret our results and extract causal links. We use a synthetic dataset $\mathbb{S}$ with samples corresponding to 500-dimensional vectors. Each vector has Gaussian noise ($\sigma = 0.4$) and a binary "pattern mask" of 5 active entries. The data set contains 30 patterns grouped into two classes. Patterns form clusters whose centroids correspond to noiseless masks. The presence of unsupervised hierarchical patterns (clusters) is a key feature of our data set. We use multilayer perceptron (MLP) networks with a single hidden layer composed of 1000 neurons with ReLU activations and no bias. Models are trained until perfect training accuracy, using SGD with a binary cross-entropy loss. All other parameters use default PyTorch values. The models in Sec. 3 are trained using the Layca optimizer [5] in full-batch mode.

2 What Noise Tells Us About Generalization

Training and Test Set Noise. Our experiments show that networks trained with moderate noise ($\sigma_{\text{train}} \approx 0.4$) generalize better across test noise levels ($0.1 \leq$

*The FRIA-FNRS funded this research through a grant attributed to Gilles Peiffer, consisting in Ph.D. funding.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

$\sigma_{\text{test}} \leq 1.2$). Low training noise prevents learning robustness, while high training noise obscures patterns, leading to memorization [6].

Internal Representations for Generalization. Fig. 1 shows the evolution of average pre-activations (in the hidden neuron with the highest absolute connection to the output [1]) for all test samples ($\sigma_{\text{test}} = 0.4$) of a pattern. The good model

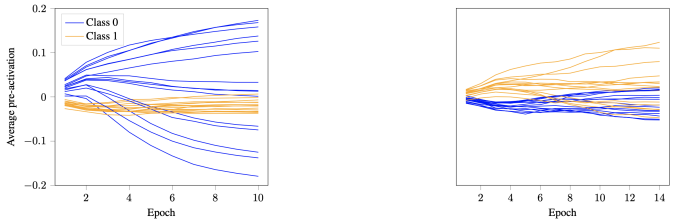

Fig. 1: Average pattern pre-activation throughout training. (Left) Good model ($\sigma_{\text{train}} = 0.4$). (Right) Bad model ($\sigma_{\text{train}} = 1.2$).

has a qualitatively different way of treating patterns than the worse model.

3 Hierarchical Patterns as Drivers of Implicit Clustering

3.1 A Winner-Take-Most Mechanism

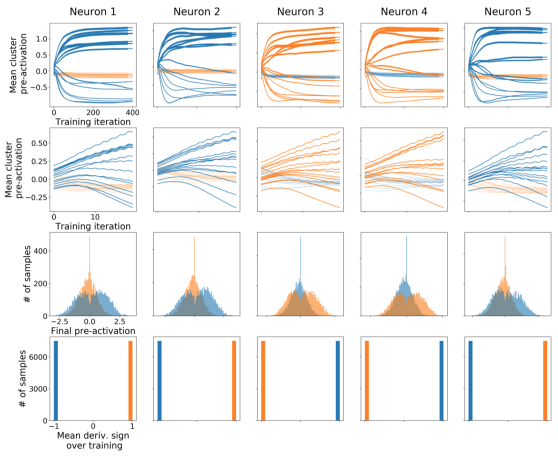

Fig. 2: (Rows 1 and 2) Evolution of average pattern pre-activations during training. (Row 3) Pre-activation distributions after training. (Row 4) Average over training of the sign of the loss derivative w.r.t. the activation.

We record two neuron-level signals for a hidden neuron $n^{(i)}$, for each sample $\boldsymbol{x}^{(j)}$, after each epoch k: (i) the pre-activation $z_{i;k}^{(j)} = (\boldsymbol{W}_{1;k}^{\top})_{i,:}\boldsymbol{x}^{(j)}$, and

[1] Chosen for illustrative purposes, as it has the most impact on the prediction of the network.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

(ii) the partial derivative of the loss w.r.t. the activation, $(\partial L/\partial a_i^{(j)})_k$, where $a_i^{(j)} = \rho(z_i^{(j)})$. After training, we select the neurons with the most influence on the output (largest absolute weights in the output layer).

During training, each neuron consistently differentiates the patterns of a class according to a winner-take-most (WTM) mechanism: patterns with larger average pre-activations are pushed towards larger pre-activations, while patterns with smaller pre-activations are pushed towards smaller pre-activations. For a given neuron, this unsupervised mechanism can be more impactful than the supervised learning process: neurons can differentiate patterns belonging to a same class more strongly than patterns from different classes.

An increased activation benefits the classification of a given sample when the average sign of the partial derivative of the loss w.r.t. the activation is negative (and inversely). This sign correlates with the sample's class across the whole training process: the samples of the class with negative mean derivative should always be pushed towards larger activations (and inversely). In practice, we see the WTM mechanism occurring with samples of the negative derivative class.

3.2 Towards Understanding the Mechanism: an Ablation Study

Fig. 3 shows several necessary elements for the WTM phenomenon: multiple hidden neurons, sufficient noise, and nonlinear activation functions. Without

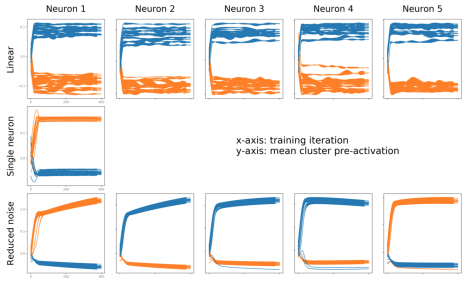

Fig. 3: Average pre-activation per cluster for a model without ReLU (row 1), with one hidden neuron (row 2), or trained on a less noisy data set (row 3). These models have poor accuracy ($\approx 80\,\%$).

them, hidden neurons classify data according to the two classes, disregarding intraclass clusters. This leads to poor accuracy ($\approx 80\,\%$).

The Role of Difficult Training Samples. Row 2 of Fig. 2 shows that around the 6th epoch, some clusters are "pushed" in the same direction as clusters of the opposite class. These are the "losing" clusters of the class subject to the WTM mechanism. This local behavior seems contrary to the global objective: differentiating examples from their opposite class. Row 4 shows the derivatives for these clusters are negative, aiming in the opposite direction they go.

This can be explained through difficult samples (whose additive noise makes them more similar to the other class). Their share of the loss increases during

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

training, as the other samples are correctly classified, and they become the dominant force driving gradient directions. Regular samples then get pushed in a direction opposed to their associated gradient, shown in Fig. 4. The difficulty

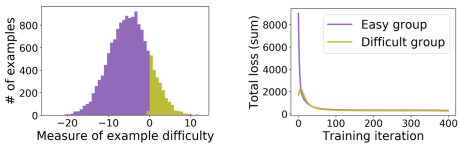

Fig. 4: (Left) The distribution of the difficulty coefficient. (Right) The evolution of the easy and difficult groups' share in the total loss during training.

coefficient for a sample $\boldsymbol{x}^{(i)}$ with class label $\mathcal{C}^{(i)}$ is $d(\boldsymbol{x}^{(i)}) = \langle \boldsymbol{x}^{(i)}, \sum_{k \notin \mathcal{C}^{(i)}} c_k \rangle - \langle \boldsymbol{x}^{(i)}, \sum_{k \in \mathcal{C}^{(i)}} c_k \rangle$. It compares the sample to the combined centroids of its class and of the other class The growing loss associated with the difficult samples leads to contradictory gradients. Around the 6th epoch, around the same time the WTM phenomenon appears in Fig. 2, the share of the difficult group becomes dominant. This aligns with our earlier conclusion: when noise is reduced, the share of difficult samples decreases, preventing the WTM mechanism.

The Role of ReLU. The overall gradient is computed as a sum of per-sample gradients, so directions that are coherent across training samples are reinforced [7, 8]. The derivative of the ReLU function is 0 for negative values, so training samples that do not activate a given neuron do not contribute to the gradient associated with that its weights. For samples presenting a certain pattern, only those that activate the neuron reinforce each other.

Samples associated with a cluster share a common pattern in the input and in the back-propagated error (because their label is the same). The number of samples that activate the neuron affects the relative share of each cluster in the gradient associated with the weights of the neuron, to which the average pre-activation of a cluster in a neuron is proportional. Difficult samples impact clusters with smaller average pre-activations more strongly, causing them to "lose" the competition. Clusters with larger average pre-activations are less affected; they "win" the competition. ReLU layers being key for differentiating "winning" and "losing" clusters is also coherent with the ablation study; the WTM mechanism requires nonlinear (e.g. ReLU) activations.

A Divide-and-Conquer Strategy. With a single hidden neuron, no WTM mechanism appears: the mechanism locally misclassifies multiple clusters, which must be countered by a correct classification in other neurons. A divide-and-conquer approach, where different neurons focus on classifying different clusters, is coherent with the WTM phenomenon. The neuron-level differences in favored clusters are determined by the random initialization of pre-activations.

Why Does the Mechanism Affect a Single Class? Jointly considering the role of difficult training samples, gradient coherence, and divide-and-conquer strategies helps explain why the WTM mechanism only applies to samples of the class

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

associated with negative derivatives, whose activations should increase during training. For this class, pushing a cluster in its "opposite" direction deactivates some of its samples and reduces the cluster's share of the neuron's gradient, promoting the correct classification of difficult samples. Inversely, pushing clusters of the positive derivative class in their opposite direction increases the number of samples that activate the neuron (and the share of the cluster in the neuron's gradient), promoting the neuron's correct classification of these clusters.

4 Related Work in Standard Deep Learning Settings

Our work reveals the emergence of a WTM mechanism and provides intuitions and experiments to understand it, but remains limited to simple data and models. To support its generality and show it still holds in standard settings, we discuss several works that studied related concepts, highlighting the links with our results. We demonstrate that our setup exhibits several phenomena occurring in standard settings and give evidence that they relate to WTM mechanisms.

Link With Training Dynamics w.r.t. Sample Difficulty. Relating notions of sample difficulty to learning speed is an active area of research. [6] showed that many samples are consistently (mis-)classified after the first training epoch, and conjectured that "deep learning learns simple patterns first, before memorizing." [9] showed that networks learn samples that are correctly classified by non-deep learning approaches first. [10] characterized samples by their consistency score (the expected accuracy as a held-out sample for varying training sizes). Their Fig. 10 shows the training curves associated with the training samples, grouped by consistency score. It reveals that samples with higher scores are learned first, and that the accuracy of low-scoring samples decreases in the first epochs of training. This suggests that the gradients of low-scoring samples "contradict" those of high-scoring samples. These results align with Fig. 4: (i) easy samples are learned before difficult ones and (ii) the loss of difficult samples increases in the first training iterations, suggesting the presence of contradictory gradients.

Link With the Coherent Gradient Hypothesis. Gradient coherence (GC) [7] plays a crucial role in the generalization abilities of deep neural networks. [8] provides experiments to support this hypothesis in standard deep learning settings, justifying the role of GC as follows: as gradients are the sum of per-sample gradients, coherence is stronger in the directions where the per-sample gradients are more similar. Training has a bias towards updates that benefit many samples, which is argued to be beneficial for generalization, based on algorithmic stability theory. However, this intuition only holds early in training, when the model misclassifies most samples. In Sec. 3.2, a small set of difficult training samples strongly influenced the overall gradient later in training. [11, 12] provide more evidence for the connection between the alignment of per-sample gradients and generalization, but more work is needed. We believe our paper offers a promising path towards a better understanding of this connection.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Link With the Benefits of Implicit Clustering Abilities. [13] proposed measures of intraclass clustering that correlate with generalization in standard deep learning settings, across variations of many hyperparameters (data augmentation, depth, learning rate, ...). Two neuron-level measures, (c_1 and c_3), capture how much samples or subclasses from the same class are differentiated in a neuron's pre-activations. They are related to the WTM mechanism, as it leads to the differentiation of clusters from the same class in a neuron's pre-activations. Our work agrees: the more the clusters of a class are differentiated, the better the generalization. Studying both implicit clustering abilities and mechanisms constitutes a coherent framework supporting the crucial role of implicit clustering.

5 Discussion and Future Work

We found that networks trained with moderate noise form selective neuronal representations through a WTM mechanism. Through an ablation study, we showed the importance of nonlinearity, representational capacity, and sufficient difficult samples. We drew links between our phenomenon and existing work on gradient coherence and intraclass clustering. Future work should investigate the effect of hyperparameters (data augmentation, pre-training, depth, learning rates) to bridge simple setups with real-world models.

References

[1] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014*.

[2] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR 2014*.

[3] Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, 2015.

[4] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020.

[5] Simon Carbonnelle and Christophe De Vleeschouwer. Layer rotation: a surprisingly simple indicator of generalization in deep networks? In *ICML 2019 Workshop*.

[6] Devansh Arpit et al. A closer look at memorization in deep networks. In *ICML 2017*.

[7] Satrajit Chatterjee. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. In *ICLR 2020*.

[8] Piotr Zielinski, Shankar Krishnan, and Satrajit Chatterjee. Explaining memorization and generalization: A large-scale study with coherent gradients. *CoRR*, 2020.

[9] Karttikeya Mangalam and Vinay Uday Prabhu. Do deep neural networks learn shallow learnable examples first? In *ICML 2019 Workshop*.

[10] Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing structural regularities of labeled data in overparameterized models. In *ICML 2021*.

[11] Satrajit Chatterjee and Piotr Zielinski. Making coherence out of nothing at all: Measuring the evolution of gradient alignment. *CoRR*, 2020.

[12] Satrajit Chatterjee and Piotr Zielinski. On the generalization mystery in deep learning. *CoRR*, 2022.

[13] Simon Carbonnelle and Christophe De Vleeschouwer. Intraclass clustering: an implicit learning ability that regularizes DNNs. In *ICLR 2021*.