

---

# Emergent representations in networks trained with the Forward-Forward algorithm

---

Niccolò Tosato<sup>1,\*</sup>Lorenzo Basile<sup>2,\*</sup>Emanuele Ballarin<sup>2</sup>Giuseppe de Alteriis<sup>3,4</sup>Alberto Cazzaniga<sup>1</sup>Alessio Ansuini<sup>1</sup><sup>1</sup>AREA Science Park, Italy<sup>2</sup>University of Trieste, Italy<sup>3</sup>King's College London, UK<sup>4</sup>University College London, UK<sup>\*</sup>Equal contributions**Correspondence:**

alessio.ansuini@areasciencepark.it

**Abstract**

The Backpropagation algorithm has often been criticised for its lack of biological realism. In an attempt to find a more biologically plausible alternative, the recently introduced *Forward-Forward* algorithm replaces the forward and backward passes of Backpropagation with two forward passes. In this work, we show that the internal representations obtained by the Forward-Forward algorithm can organise into category-specific *ensembles* exhibiting high sparsity – i.e. composed of an extremely low number of active units. This situation is reminiscent of what has been observed in cortical sensory areas, where neuronal ensembles are suggested to serve as the functional building blocks for perception and action. Interestingly, while this sparse pattern does not typically arise in models trained with standard Backpropagation, it can emerge in networks trained with Backpropagation on the same objective proposed for the Forward-Forward algorithm. These results suggest that the learning procedure proposed by Forward-Forward may be superior to Backpropagation in modelling learning in the cortex, even when a backward pass is used.

## 1 Introduction

Deep Learning is a highly effective approach to artificial intelligence, with tremendous implications for science, technology, culture, and society [1]. At its core, there is the Backpropagation (Backprop) algorithm [2], which efficiently computes the gradients necessary to optimise the learnable parameters of an artificial neural network. Backprop, however, lacks biological plausibility [3] – leading to many attempts to address the issue. The most recent of such, the Forward-Forward algorithm [4], eliminates the need to store neural activities and propagate error derivatives along the network. In standard classification context, the application of Forward-Forward requires the designation of positive and negative data. E.g., to classify images, one could assign positive (or negative) data to those images having their correct (or incorrect, respectively) class embedded via one-hot encoding at the border. The Forward-Forward algorithm then learns to discriminate between positive and negative data by optimising a goodness function (e.g., the  $\ell_2$  norm of the activations), akin to contrastive learning [5]. Satisfactory results have been observed [4] for classification tasks on MNIST [6], a well-known benchmark dataset.

This work goes beyond mere performance, investigating the presence of similarities between biological and artificial neuronal ensembles. Our experiments demonstrate sparsity in representations learned using Forward-Forward, and similarities between these and cortical ensembles found in early stages of sensory processing [7]. Neurons that form ensembles are highly specialised, and are found to activate selectively when presented with positive data, while only exhibiting minimal activation to negative data or unstructured stimuli. Our results also indicate that in such representations there may exist collective suppression mechanisms similar to those of biological inhibitory neurons [8]. Finally, though optimising the cross-entropy loss for the same classification task does not appear to produce the sparse ensembles we observe, the phenomenon may not solely be due to the use of the Forward-Forward algorithm. In fact, similar results are obtained by optimising the same goodness function of Forward-Forward, with Backprop instead. This suggests that more focus should be put on the purpose and biological meaning of the loss function rather than the training algorithm [9].

## 2 Related work

In the section that follows, we summarise key aspects of the Forward-Forward algorithm and the main findings pertaining identification and characterisation of biological neuronal ensembles in the cortex.

### 2.1 Forward-Forward

The Forward-Forward algorithm [4] is a newly proposed learning algorithm for artificial neural networks, whose main premise is the ability to overcome the notorious biological implausibility of Backprop [2]. In fact, while the effectiveness of Backprop makes it the standard algorithm for training neural networks, it is based on biologically unrealistic assumptions, such as the need to propagate information forwards and backwards through the network [9].

Forward-Forward owes its name to the fact that it replaces the backward pass with an additional forward pass. The two forward passes are executed on different data, named positive and negative data. During training, the objective of Forward-Forward is to maximise a so-called goodness function of the neural activations (e.g. the  $\ell_p$  norm) on positive data and minimise it on negative data. In a simple image classification setting, such as the one we adopt in this paper, one could encode a class label at the border of images, by one-hot encoding it with a white pixel (as shown in Figure 1, Panel A). Then, following the definition from [4], positive data are those for which the encoded label matches the ground truth label, while the opposite holds for negative data.

Layers are trained separately and sequentially, and learn to discriminate between positive and negative data by maximising and minimising their goodness, according to the data presented. Crucially, activations are normalised before being passed to the subsequent layer, to prevent layers from relying on the goodness computed by their predecessors. From the biological point of view, normalisation is known to be a “canonical neural computation” [10]. At test time, when a new unlabelled sample has to be categorised, many copies of the image are created, each with a different one-hot encoded label. These are then fed into the neural network to obtain a goodness score. Finally, the image gets classified in the category that produced the maximum goodness value.

In the seminal Forward-Forward paper [4], satisfactory classification results are reported on the standard handwritten digit recognition dataset MNIST, with the definition of positive and negative data described above, and using the  $\ell_2$  or  $\ell_1$  norm of activations as goodness function.

### 2.2 Neuronal ensembles

In Neuroscience, ensembles are defined as sparse groups of neurons that co-activate during spontaneous activity or in response to sensory stimuli. Remarkably, ensembles – rather than single neurons – have long been suggested to be emergent functional blocks of cortical activity [8, 11, 12, 13] and have a prominent role in sensory processing, memory [14] and behaviour [15]. As shown by [7], during visual processing the spiking activity in the cortex is dominated by ensembles, whose properties cannot be accounted for by the independent firing activity of neurons in isolation. Additionally, ensembles can be both generated by sensory stimuli (visual, in this case) or by the spontaneous activity of the network; however, visually evoked ones are time-locked to stimuli. It is also shown that single neurons can participate in more than one ensemble, thus maximising the encoding potential of the network. Interestingly, the same ensembles are evoked both when reacting to stimuli and during

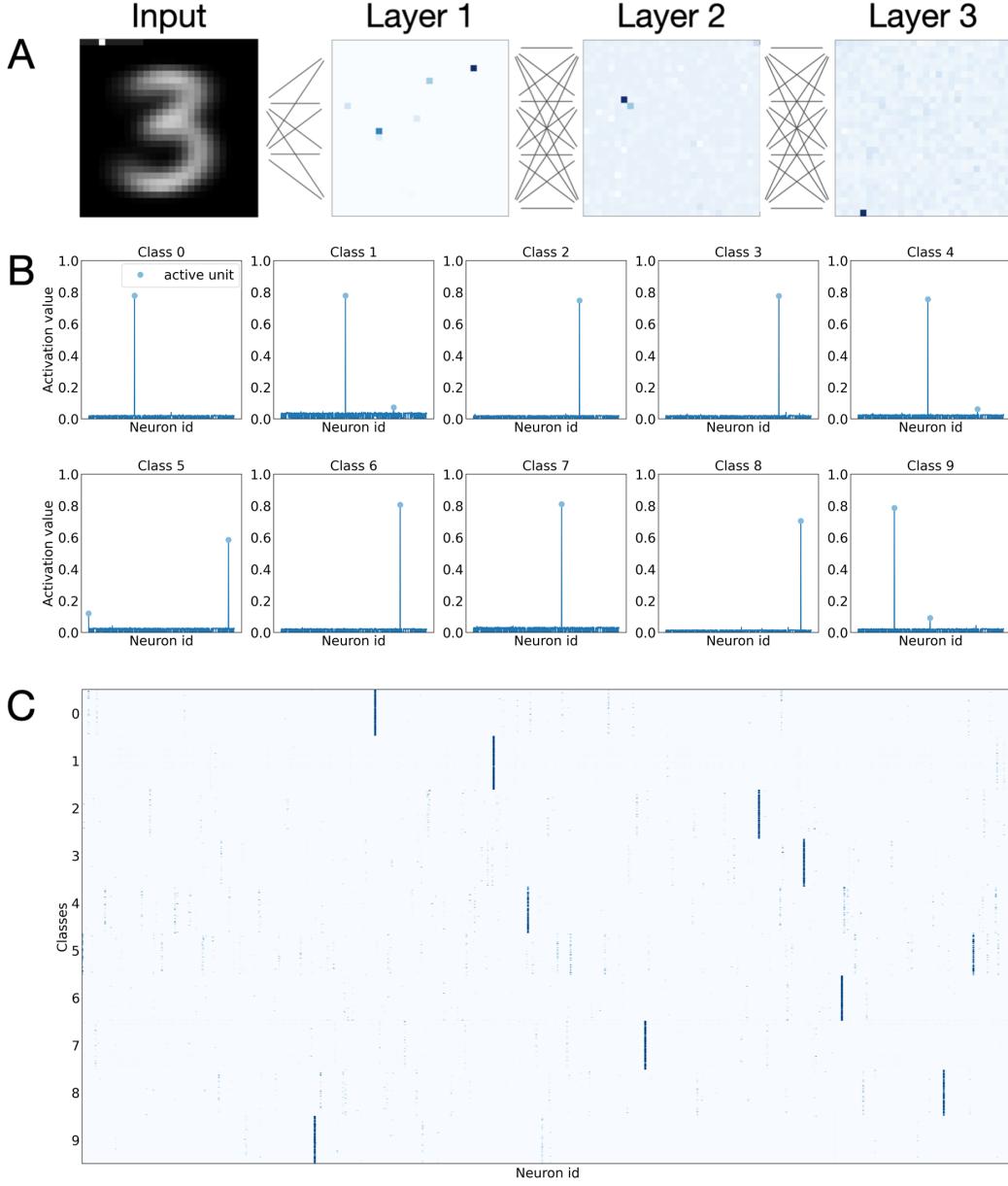


Figure 1: Activation patterns in a Multi-Layer Perceptron trained with the Forward-Forward algorithm, on the MNIST dataset.

Panel A Examples of activation patterns in response to a typical input. Images show the activation value for network units, arranged as a square for the sake of clarity; darker squares represent more active neurons.

Panel B Activation value of each neuron in the first hidden layer (Layer 1), averaged on all images of a given class. Neuron index on the  $x$  axis; average activation on the  $y$  axis. Blue dots indicate units that are considered active according to the method described in subsection 3.5.

Panel C Activation map for neurons in Layer 1 for all images, grouped by class. A blue dot in position  $(x, y)$  indicates that neuron  $x$  is activated by input  $y$ ; colorscale represents the intensity of such activation. Horizontal bands mark different categories; bright vertical lines mark active neurons. Each input category activates consistently a specific sets of neurons (ensemble).

spontaneous activity, indicating that they may serve as inherent fundamental components giving

rise to visual responses. Findings from [16] confirm that visual inputs activate sparse ensembles in the primary visual cortex (V1). Moreover, images can be decoded starting from a low number of highly responsive neurons (i.e. a highly sparse ensemble), and adding more does not increase (or even harms) decoding performance. These neurons consistently and reliably encode the same visual inputs across multiple trials. For that reason, partially overlapping receptive fields may play a role in achieving sparse and robust representation of information. Consequently, collecting the activity of highly responsive neurons emerges as an optimal decoding strategy for downstream neurons. Ensembles are also shown to be present in other animal models [17, 18] and recent works propose that they may also play a role in conscious experience [19]. The perdurance of ensembles – they have the capacity to endure for weeks [20] – suggests that they could also potentially function as a foundation for the long-term representation of perceptual states or memories.

Advancements in technology have allowed not only to visualise, but also to stimulate cortical neurons [21], allowing the execution of complex tasks from ensembles of neurons [22]. These approaches show [15] that when a specific group of neurons in the V1 of mice is repeatedly activated, it generates a coactive (i.e. imprinted) neuronal ensemble that remains spontaneously active even after a relatively long time has passed. Remarkably, the activation of certain individual neurons from these imprinted ensembles is capable of re-evoking the entire ensemble, showcasing a phenomenon of pattern completion. This effect persists long after the initial imprinting process. [22] also show a causal relation between stimulation of ensembles and behaviour.

Finally, the idea of neuronal ensembles has inspired the development of computational models [23]. E.g., [24] shows that sparse and redundant (hence more robust) representations are optimal for encoding natural images with noisy or unreliable neurons, a finding as also confirmed by [25] and [26].

### 3 Methods

In this work, we investigate and compare the representations produced by three models:

- A classifier in the style of that used by [4], trained with Forward-Forward (**FF**);
- A classifier identical to the above, but trained end-to-end with Backprop to optimise the same goodness function (**BP/FF**);
- A classifier trained with Backprop on the categorical cross-entropy loss, as customary (**BP**).

Such different scenarios are further analysed individually in subsection 3.2, subsection 3.3 and subsection 3.4, respectively.

#### 3.1 Data

The datasets we use to train and test the models described so far are MNIST [6], FASHIONMNIST [27] and SVHN [28]. Details on these datasets are provided in the Appendix.

#### 3.2 Model trained with Forward-Forward (**FF**)

Our **FF** model is inspired by the architecture proposed by [4] – and likewise trained according to the Forward-Forward algorithm. It consists of fully-connected layers (3 for MNIST and FASHIONMNIST, 4 for SVHN), each composed by a number of units equal to the input dimension (784 for MNIST and FASHIONMNIST, 1024 for SVHN), with element-wise sigmoidal activations. Both during training and inference, the layer-wise  $\ell_\infty$  norm is used as the goodness function of choice. Correspondingly,  $\ell_\infty$  normalisation is performed between subsequent layers.

To define positive and negative data, a one-hot-encoded class vector is embedded at the top-left corner of images. Prior to such embedding, these pixels are set to black colour. Then, in the case of positive data, the pixel corresponding to the true class is switched to the maximum value elsewhere observed in the image, while in the case of negative examples such value is randomly assigned to one of the other pixels of the embedding vector. In the case of RGB datasets (such as SVHN), the class embedding is replicated identically on each channel.

### 3.3 Model trained with Backpropagation on the goodness objective (BP/FF)

The architecture of the **FF** model, while designed to be optimised using the Forward-Forward algorithm, can be trained seamlessly with Backpropagation on the same goodness maximisation/minimisation objective. Indeed, keeping the definition of positive and negative data introduced for **FF**, one could simply use Backprop to optimise the goodness-based loss from the Forward-Forward algorithm.

In detail, positive and negative data are fed to the network during the forward step, and the overall goodness of the internal representation is evaluated. The backward pass is then executed, and parameters are optimised to achieve the same goal as the **FF** model. It is worth pointing out that, in this case, the goodness is maximised globally instead of layer-by-layer (i.e., locally).

### 3.4 Model trained with Backpropagation on the cross-entropy loss (BP)

The **FF** and **BP/FF** models are also compared to a standard neural classifier, serving as a baseline. For such purpose, a multi-layer perceptron is employed. The model shares the same number of layers, layerwise neuron count, and non-linear activation function choice with **FF** and **BP/FF**. The only architectural difference between the **BP** model and the other two is the addition of a final softmax layer, to suitably shape and scale the output for the classification task. The model is trained end-to-end with Backprop on the categorical cross-entropy loss.

### 3.5 Analysis of representations

For each model described, we analyse the internal representation emerging at each layer. We limit our analysis to data belonging to the test set (i.e. not seen during training) and correctly classified by the respective model.

We first assess the layer-wise sparsity of representations by computing their mean absolute deviation from the mean (MAD). For each layer we extract a representation matrix  $X$  of dimension  $(M, n)$ , where  $M$  is the total number of test images and  $n$  is the number of neurons in the layer considered.

We then compute the MAD of the representation as  $\frac{\sum_{ij} |X_{ij} - \mu|}{Mn}$ , where  $\mu$  is the mean of  $X$ . The rationale for choosing the MAD as an indicator of sparsity is that 1) representations are sparse when only a few neurons activate in response to each data sample, and 2) the activation level is significant only when it departs from a baseline, which we set as the average activation value  $\mu$ . In Table 2 we report the MAD values for all our experiments. In the Appendix, we also report results relative to a different useful metric, which quantifies the skewness of representations.

To detect the emergence of ensembles, within each model and dataset combination, we adopt the following method.

We start by defining a category-specific representation matrix  $X_c$ , of shape  $(M_c, n)$ , where  $M_c$  is the number of correctly classified test images of the given category and  $n$  is the number of neurons in the layer considered. Then, we compute the histogram of values in  $X_c$ , set a quantile in their distribution (also called *threshold* hereinafter), and define a neuron as active – and therefore part of the evoked ensemble – if its median activation (across columns of  $X_c$ ) exceeds the threshold. Increased robustness to noisy observations justifies the choice of the median instead of the mean.

The output of the ensemble computation is a set of active units for each category:  $\mathcal{E}^c = \{e_1^c, e_2^c, \dots, e_{n_c}^c\}$ ,

$\forall c \in \{1, 2, \dots, C\}$ , where  $n_c$  is the number of active units for category  $c$ . Once the ensembles are defined, it is possible to look at units that are shared across categories  $c$  and  $c'$  by considering  $\mathcal{E}^c \cap \mathcal{E}^{c'}$ . Examples of shared units are reported in Figure 2 and Figure 3.

When the representation is not sparse (a high MAD value is indicative of this) ensembles are ill-defined as too many neurons significantly active simultaneously. In these cases, we do not compute ensembles.

## 4 Results

We describe our findings for the three models introduced, on the MNIST, FASHIONMNIST and SVHN datasets. In particular, we focus on the unique properties of representations obtained within the **FF** model, i.e. a model trained with Forward-Forward on its natural goodness objective. Such properties – e.g. the emergence of category-specific ensembles and the presence of shared units across them –

establish a link between neural networks trained with the Forward-Forward algorithm and biological cortical networks described in subsection 2.2.

#### 4.1 Classification accuracy

Before delving into the main results of this work, we evaluate the performances of our models on the classification tasks at hand. Table 1 contains results in terms of test set classification accuracy for all models we employed – **FF**, **BP/FF** and **BP** – on MNIST, FASHIONMNIST and SVHN. While some of these accuracy values are far from the state-of-the-art (i.e., respectively, 0.997 [29], 0.931 [27] and 0.860 [30], for fully-connected networks), they are a solid ground on which to build our subsequent investigations. Training details and hyperparameters for all models are reported in the Appendix.

Table 1: Test-set classification accuracy for the models considered in our investigation. Results expressed as *mean*  $\pm$  *std. dev.* over 10 runs with independent randomised weight initialisation.

Dataset	<b>FF</b>	<b>BP/FF</b>	<b>BP</b>
MNIST	$0.940 \pm 0.004$	$0.935 \pm 0.025$	$0.984 \pm 0.001$
FASHIONMNIST	$0.826 \pm 0.007$	$0.822 \pm 0.018$	$0.887 \pm 0.005$
SVHN	$0.618 \pm 0.003$	$0.805 \pm 0.004$	$0.851 \pm 0.002$

#### 4.2 Forward-Forward elicits sparse neuronal ensembles

Table 2: Mean absolute deviation from the mean of representations values. Results expressed as *mean*  $\pm$  *std. dev.* over 10 runs with independent randomised weight initialisation.

Dataset	Layer	<b>FF</b>	<b>BP/FF</b>	<b>BP</b>
MNIST	1	$0.015 \pm 2.1 \cdot 10^{-4}$	$0.004 \pm 2.1 \cdot 10^{-4}$	$0.394 \pm 2.9 \cdot 10^{-3}$
	2	$0.079 \pm 9.8 \cdot 10^{-4}$	$0.039 \pm 1.7 \cdot 10^{-2}$	$0.254 \pm 2.5 \cdot 10^{-3}$
	3	$0.005 \pm 3.3 \cdot 10^{-4}$	$0.003 \pm 1.1 \cdot 10^{-4}$	$0.392 \pm 1.3 \cdot 10^{-3}$
FASHIONMNIST	1	$0.022 \pm 5.2 \cdot 10^{-4}$	$0.004 \pm 4.6 \cdot 10^{-4}$	$0.414 \pm 4.3 \cdot 10^{-3}$
	2	$0.120 \pm 1.7 \cdot 10^{-3}$	$0.065 \pm 3.7 \cdot 10^{-2}$	$0.234 \pm 5.4 \cdot 10^{-3}$
	3	$0.113 \pm 3.9 \cdot 10^{-3}$	$0.003 \pm 1.1 \cdot 10^{-3}$	$0.379 \pm 5.9 \cdot 10^{-3}$
SVHN	1	$0.019 \pm 2.1 \cdot 10^{-3}$	$0.452 \pm 2.9 \cdot 10^{-3}$	$0.463 \pm 1.3 \cdot 10^{-3}$
	2	$0.086 \pm 9.4 \cdot 10^{-4}$	$0.010 \pm 9.3 \cdot 10^{-4}$	$0.242 \pm 4.7 \cdot 10^{-2}$
	3	$0.036 \pm 1.4 \cdot 10^{-3}$	$0.107 \pm 2.0 \cdot 10^{-2}$	$0.145 \pm 7.1 \cdot 10^{-3}$
	4	$0.098 \pm 3.2 \cdot 10^{-2}$	$0.014 \pm 2.1 \cdot 10^{-3}$	$0.245 \pm 9.2 \cdot 10^{-3}$

The **FF** and **BP/FF** models – based on the original Forward-Forward network architecture, and trained according to the goodness objective (subsection 3.2 and subsection 3.3) – exhibit the co-activation of small neuronal ensembles, similar to those observed in cortical representations [8, 7, 12].

Figure 1 (Panels **B**, **C**) shows an example of neuron activations in Layer 1 of the **FF** model trained on MNIST, and showcases the emergence of sparse, category-specific, ensembles (see the Appendix for a similar visualisation for Layers 2 and 3 of the same model). These representations can involve a remarkably low number of active units: ensembles composed by just one or a few neurons are often observed, also when the **FF** model is trained on the more challenging FASHIONMNIST and SVHN datasets (Table 3).

The emergence of small sized ensembles is evident also in representations from the **BP/FF** model – although the average number of active units per ensemble is larger than for **FF** (Table 3).

Sparse representations, as is the case for **FF** and **BP/FF**, involve low MAD values (Table 2), and in this case well defined ensembles typically emerge. On the contrary, representations in the **BP** model are accompanied by large MAD values, caused by a large fraction of neurons that simultaneously activate on each data sample. In these case the ensembles are ill-defined.

Table 3: Average ensemble sizes. Ensembles are defined according to the method presented in subsection 3.5, with threshold set at 0.98. Results expressed as  $mean \pm std. dev.$  over 10 runs with independent randomised weight initialisation.

Dataset	Layer	FF	BP/FF
MNIST	1	$1.58 \pm 0.26$	$2.50 \pm 0.46$
	2	$1.02 \pm 0.04$	$3.31 \pm 0.86$
	3	$1.00 \pm 0.00$	$1.55 \pm 0.55$
FASHIONMNIST	1	$2.19 \pm 0.17$	$4.13 \pm 1.06$
	2	$1.20 \pm 0.11$	$4.83 \pm 2.53$
	3	$1.16 \pm 0.14$	$1.70 \pm 0.58$
SVHN	1	$2.34 \pm 0.25$	not sparse
	2	$6.22 \pm 0.17$	$18.28 \pm 6.42$
	3	$6.50 \pm 0.33$	$10.73 \pm 0.47$
	4	$10.20 \pm 0.81$	$7.69 \pm 1.15$

### 4.3 Semantically similar classes elicit ensembles with shared neurons

Drawing a parallel with a phenomenon observed in Neuroscience [16], related categories can be expected to share units of their ensembles. This is indeed what we observe, as shown in Figure 2. Results are reported for FASHIONMNIST, where different classes of clothes or shoes may contain a common share of visual features. In this regard, we observe a clear tendency to share units between similar classes – e.g. across representations of pullover, coat and shirt or sneaker and ankle boot.

In the following section, we provide evidence that a unit can be shared across two ensembles even if one of these refers to an unseen category (i.e., excluded from the training set but whose representation, extracted at test time, generates a valid ensemble), as we show in Figure 3 (Panel C).

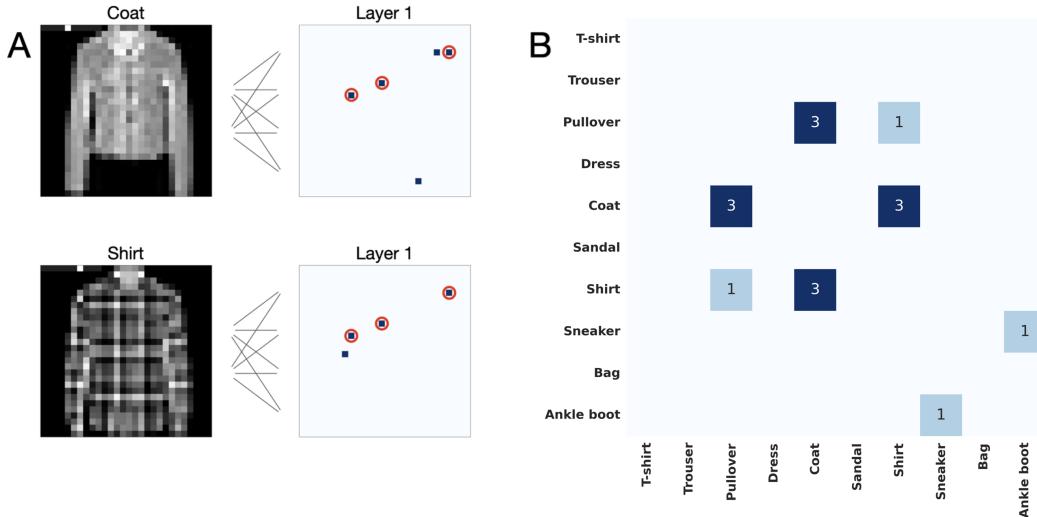


Figure 2: Semantically similar classes in FASHIONMNIST can elicit ensembles with shared neurons. Panel A The ensembles elicited in the first hidden layer by two example inputs. Red circles indicate the active units which are shared between the two categories.

Panel B Element  $i, j$  of the matrix indicates how many units are shared between the ensembles of category  $i$  and category  $j$ , with reference to a single training run.

#### 4.4 Representations of unseen categories can contain ensembles

We investigate the ability of a trained **FF** model to respond to unseen categories with a coherent activation pattern which is typical of the ensembles we found on the categories seen at training time. To this end, we repeatedly train **FF** on FASHIONMNIST, removing one category at a time. Then, we extract the representation of the missing category, and verify if a valid ensemble is formed. We find that, in six out of ten cases (see Figure 3 for one example, and the Appendix for a more detailed account), this is indeed the case, with the new ensembles sharing the same characteristics as the ones emerging for seen categories, apparently with the only exception of a lower average activation of their constituent neurons.

In several cases, we also found that the ensembles of unseen categories share units with the ensembles of seen categories, when endowed with similar visual features or semantics (Figure 3, Panel **C**). A more extensive exploration of these cases is reported in the Appendix.

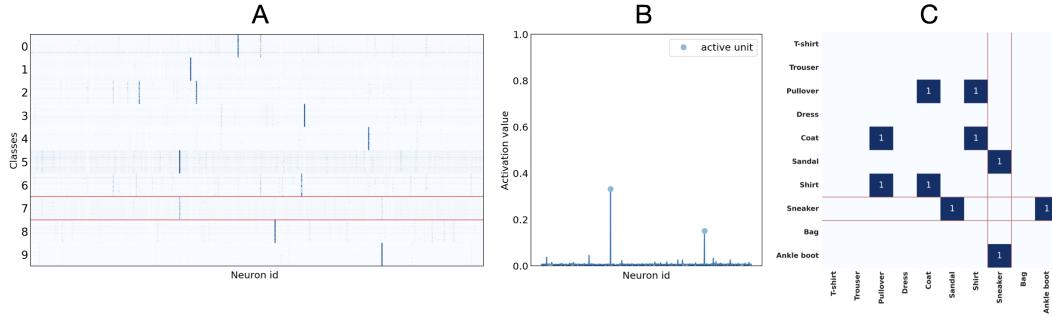


Figure 3: The representation of an unseen category forms a valid ensemble in **FF** trained on FASHIONMNIST.

Panel A Activation patterns in response to the different categories in the first hidden layer. The unseen category (**Sneakers**) is surrounded by red lines.

Panel B Activation value of each neuron, averaged on all images of the unseen category. Neuron index on the  $x$  axis; average activation on the  $y$  axis. Blue dots indicate units that are considered active according to the method described in subsection 3.5.

Panel C Ensembles of unseen categories can share units with the ensembles of the other categories.

#### 4.5 Distribution of excitatory and inhibitory connections

We observed in subsection 4.2 that the size of ensembles in **BP/FF** is just slightly larger than in **FF** (Table 3). The emergence of such sparse ensembles suggests that a large fraction of neurons in those two models should be strongly inhibited. Therefore, a natural question to ask in our setting is: “what is the fraction of positive (i.e. excitatory) weights w.r.t. the total, in the different layers of the models we study?”

To this end, we consider for each neuron  $i$  in a layer with  $n$  neurons the fraction of its positive weights w.r.t. the total number of its afferent connections ( $\varrho_i^+$  in the following), together with the layer average  $\varrho^+ = \frac{1}{n} \sum_i \varrho_i^+$ . A neuron is strongly imbalanced towards inhibition when  $\varrho_i^+ \approx 0$  and, viceversa, strongly imbalanced towards excitation when  $\varrho_i^+ \approx 1$ ; when its  $\varrho_i^+ \approx 0.5$  we will say that the neuron is almost perfectly balanced.

Focusing on the second hidden layer, we observe macroscopic differences of the empirical distribution of  $\varrho_i^+$  among all the three models (Figure 4), with 1) a marked dominance of inhibition – as expected – in the case of **FF**, 2) a bimodal distribution of  $\varrho_i^+$  in **BP/FF** revealing a subpopulation of strongly inhibited neurons and a subpopulation of strongly excited ones and 3) a situation of approximate balance for all the neurons in the **BP** model. For an account of the average values  $\varrho^+$  for all models, layers and datasets used we refer to the Appendix.

From these findings, we can conclude that not only the training objective (i.e. goodness-based vs. categorical cross-entropy minimisation), but the specific training protocols are crucial to give rise to a different interplay between excitation and inhibition.

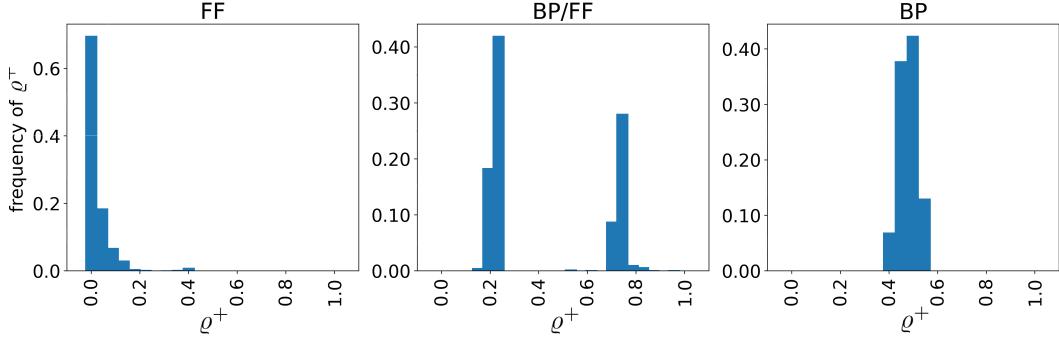


Figure 4: Distribution of the fraction of positive (i.e. excitatory) weights for each neuron in Layer 2 of the **FF**, **BP/FF** and **BP** models trained on the MNIST dataset. (Left) In **FF**, the distribution is strongly imbalanced, with around 70% of the neurons having 100% inhibitory weights. (Center) In **BP/FF** the distribution is markedly bimodal and slightly imbalanced towards inhibitory connections. (Right) The **BP** model is almost perfectly balanced between excitation and inhibition.

## 5 Discussion

The main finding of our work is that artificial neural networks – trained with the Forward-Forward algorithm – can elicit biologically plausible representations in the form of sparse neuronal ensembles [8, 7], which robustly and reliably encode meaningful information. Such ensembles can also share units across semantically related inputs [16].

We started our investigation by collecting and analysing representations from **FF** networks trained on MNIST, FASHIONMNIST and SVHN and defined, separately for each category, subsets of units (*ensembles*) that prominently and consistently activate in response to data in such category (subsection 4.2). These category-specific ensembles turn out to be composed of a few units, which is consistent with experimental findings on the sensory cortex [7]. Furthermore, when image categories are characterised by a certain degree of visual similarity, the corresponding ensembles often share one or more units (Figure 2, Panel **B**). These results are consistent with phenomena observed in sensory cortices [16] – as discussed in detail in subsection 2.2 – where ensembles are defined.

We then tested the ability of trained **FF** models to cope with new data, and we observed that the activations in response to an unseen input category formed in many cases a new ensemble, with similar characteristics of sparsity as the ones formed for other classes during training (Figure 3, Panels **A** and **B**). Furthermore, we noticed that the ensembles of unseen categories can share units with the ensembles of categories used as training data (Figure 3, Panel **C**). These findings suggests that **FF** can perform well in zero-shot classification tasks, which is particularly relevant in view of the importance of zero/few-shot learning in human and animal cognitive performances [31].

While absent in the **BP** model, the existence of ensembles composed by a few units is not unique to **FF**. It is indeed observed also in **BP/FF** (subsection 4.2), where the ensembles turns out to be slightly larger. However, despite their similarity, the **FF** and **BP/FF** models are profoundly different, as demonstrated by the different interplay between inhibition and excitation in these models (Figure 4). We observe in this regard that the excitatory/inhibitory (E/I) balance plays a key role in the stability of cortical networks and in brain dynamics [32, 33].

The sparsity of representations has computational benefits for sensory processing. [26] underline that sparsity may be the optimal encoding strategy for neuronal networks, for it is energy efficient. Sparsity also increases the memory-storage capacity, and eases readout at subsequent processing layers. [34] show that sparse and expansive coding (i.e., from a lower dimensional sensory input space to a higher dimensional neural representation) reduces the intra-stimulus variability, maximises the inter-stimulus variability, and allows optimal and efficient readout of downstream neurons. This is the reason why sparse and expansive transformations are widespread in biology, e.g., in rodents [35] or flies [36].

Overall, our findings – focused on the emerging properties of representations – corroborate the idea that Forward-Forward might be a better model than Backprop for learning in the cortex [4]. As shown in section 4, besides being a non biological learning rule, Backprop elicits non-sparse and less biologically plausible representations.

This work is a starting point for further explorations in the field of biologically plausible representations. Promising avenues for future research in this field encompass exploring model compression through pruning [37], with the design of new strategies based on the relevance of the ensembles, as well as delving into the dynamic evolution of ensemble size and organisation throughout the training process. Such investigations hold the potential to shed light on the formation, evolving interactions, and persistence or replacement of ensembles. Furthermore, these findings can be cross-referenced with neurophysiological data to assess the biological plausibility of ensembles in dynamic learning contexts.

## Acknowledgments and Disclosure of Funding

The authors thank Alex Rodriguez (University of Trieste) for the idea of testing the models on a category unseen during training.

The authors acknowledge the AREA Science Park supercomputing platform ORFEO made available for conducting the research reported in this paper and the technical support of the Laboratory of Data Engineering staff. N.T. was supported by the Italian PNR grant “FAIR-by-design”; A.C. and A.A. were supported by the ARGO funding program.

## References

- [1] Tyna Eloundou, Sam Manning, Pamela Mishkin and Daniel Rock. Gpts are gpts: an early look at the labor market impact potential of large language models, 2023. arXiv: 2303.10130.
- [2] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
- [3] David G. Stork. Is backpropagation biologically plausible? In *Proceedings of the International Joint Conference on Neural Networks*, 1989.
- [4] Geoffrey Hinton. The forward-forward algorithm: some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020.
- [6] Yann LeCun and Corinna Cortes. The MNIST handwritten digit database, 2010.
- [7] Jae-eun Kang Miller, Inbal Ayzenshtat, Luis Carrillo-Reid and Rafael Yuste. Visual stimuli recruit intrinsically generated cortical ensembles. *Proceedings of the National Academy of Sciences*, 111(38):E4053–E4061, 2014.
- [8] Rafael Yuste. From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16(8):487–497, 2015.
- [9] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019.
- [10] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2011.
- [11] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.
- [12] Kenneth D Harris. Neural signatures of cell assembly organization. *Nature reviews neuroscience*, 6(5):399–407, 2005.
- [13] Buzsáki György. Neural syntax: cell assemblies, synapsemes, and readers. *Neuron*, 68(3):362–385, 2010.
- [14] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- [15] Luis Carrillo-Reid, Shuting Han, Weijian Yang, Alejandro Akrouh and Rafael Yuste. Controlling visually guided behavior by holographic recalling of cortical ensembles. *Cell*, 178(2):447–457, 2019.
- [16] Takashi Yoshida and Kenichi Ohki. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature communications*, 11(1):872, 2020.

- [17] Christophe Dupre and Rafael Yuste. Non-overlapping neural networks in hydra vulgaris. *Current Biology*, 27(8):1085–1097, 2017.
- [18] Jing Liu and Scott C Baraban. Network properties revealed during multi-scale calcium imaging of seizure activity in zebrafish. *Eneuro*, 6(1), 2019.
- [19] Richard Boyce, Robin F Dard and Rosa Cossart. Cortical neuronal assemblies coordinate with eeg microstate dynamics during resting wakefulness. *Cell Reports*, 42(2), 2023.
- [20] Jesús E. Pérez-Ortega, Tzitzitlini Alejandro-García and Rafael Yuste. Long-term stability of cortical ensembles. *Elife*, 10:e64449, 2021.
- [21] Adam M Packer, Lloyd E Russell, Henry WP Dalgleish and Michael Häusser. Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nature methods*, 12(2):140–146, 2015.
- [22] Luis Carrillo-Reid and Rafael Yuste. Playing the piano with the cortex: role of neuronal ensembles and pattern completion in perception and behavior. *Current opinion in neurobiology*, 64:89–95, 2020.
- [23] Christos H Papadimitriou, Santosh S Vempala, Daniel Mitropolsky, Michael Collins and Wolfgang Maass. Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences*, 117(25):14464–14472, 2020.
- [24] Eizaburo Doi and Michael Lewicki. Sparse coding of natural images using an overcomplete set of limited capacity units. *Advances in Neural Information Processing Systems*, 17, 2004.
- [25] David J Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, 1994.
- [26] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004.
- [27] Han Xiao, Kashif Rasul and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. arXiv: 1708.07747.
- [28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *Deep Learning and Unsupervised Feature Learning Workshop, NeurIPS*, 2011.
- [29] Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella and Jürgen Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, 22(12):3207–3220, December 2010.
- [30] Stamatios Pitsios. SVHN number recognition using deep learning. <https://github.com/pitsios-s/SVHN>, 2017.
- [31] Brenden M Lake, Ruslan Salakhutdinov and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- [32] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.
- [33] Gustavo Deco, Adrián Ponce-Alvarez, Patric Hagmann, Gian Luca Romani, Dante Mantini and Maurizio Corbetta. How local excitation–inhibition ratio impacts the whole brain dynamics. *Journal of Neuroscience*, 34(23):7886–7898, 2014.
- [34] Baktash Babadi and Haim Sompolinsky. Sparseness and expansion in sensory representations. *Neuron*, 83(5):1213–1226, 2014.
- [35] Peter Mombaerts, Fan Wang, Catherine Dulac, Steve K Chao, Adriana Nemes, Monica Mendelsohn, James Edmondson and Richard Axel. Visualizing an olfactory sensory map. *Cell*, 87(4):675–686, 1996.
- [36] Glenn C Turner, Maxim Bazhenov and Gilles Laurent. Olfactory representations by drosophila mushroom body neurons. *Journal of neurophysiology*, 99(2):734–746, 2008.
- [37] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020.
- [38] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2017.

## A Computational resources

Training and subsequent experiments were conducted on an NVIDIA DGX A100 system. The system is equipped with 8 NVIDIA A100 GPUs, interconnected by NVLink technology. Each GPU is equipped with 6912 CUDA cores, 432 Tensor cores and 40 GB of high-bandwidth memory.

## B Data

The MNIST dataset consists of pictures of handwritten Arabic numerals, from 0 to 9, each represented as a grayscale image of size  $28 \times 28$ . FASHIONMNIST has been designed as a drop-in replacement to MNIST, offering a more challenging classification task. It consists of ten classes of clothing items, still represented as grayscale images with a resolution of  $28 \times 28$ . Both datasets provide 60000 training and 10000 test images, balanced in terms of per-class numerosity.

SVHN contains coloured images of digits from house numbers, captured by Google StreetView. The images are composed of  $32 \times 32$  RGB-encoded pixels. This dataset is slightly larger than the previous two, as it contains 73257 data-points in the training set and 26032 in the test set.

The SVHN images have been cropped in order to center the digit of interest within the frame. However, the presence of adjacent digits and other distracting elements, that have been kept within the images, introduces an additional layer of complexity when compared to MNIST and FASHIONMNIST, where the subjects are prominently displayed against a uniform black background.

## C Training details

All our models (**FF**, **BP/FF** and **BP**), on all datasets (MNIST, FASHIONMNIST and SVHN), have been optimised using Adam [38]. **BP/FF** and **BP** models were trained using batches of size 2048 and a learning rate of  $10^{-3}$ , for respectively 1000 and 100 epochs. **FF** models were trained on a full-batch for 5000 epochs per layer, with the exception of the model trained on SVHN, whose first layer was trained for 20000 epochs.

For all models trained on SVHN we employed Dropout ( $p = 0.2$ ) to favour generalisation and prevent overfitting. We stress that while the choice of using Dropout and full-batch optimisation turns out to be convenient in terms of generalisation, they are by no means required to elicit sparse representations or emergence of ensembles.

## D Activation patterns in deeper layers

In subsection 4.2 we claimed that in **FF** and **BP/FF** the images of a given category activate consistently a small set of units that we named ensembles, that share similarities to what is observed in sensory cortices. We reported in Figure 1 the activation map for Layer 1 (the first hidden layer) of **FF** trained on the MNIST dataset, and observed that very sparse ensembles emerge. In this section we show, in a similar fashion, the representations for Layers 2 and 3 (Figure 5 and Figure 6, respectively). We found that an extreme sparsity is preserved across all layers of the network; a similar scenario emerges for similar models trained on FASHIONMNIST, SVHN and also for **BP/FF** on all datasets (Table 3).

## E Skewness of representations

Expanding on the findings presented in Table 2 regarding the sparsity of representations learned by **FF** and **BP/FF**, we present here an additional analysis of the distribution of activations. Table 4 reports the average Fisher-Pearson moment coefficient of skewness, over 10 runs with independent random weight initialisation, for all models and datasets. It is evident that the representations obtained by **FF** and **BP/FF** are very positively skewed, meaning that they contain activations that greatly deviate from the mean (the neurons that form ensembles), with the only exception of Layer 1 of the **BP/FF** classifier on SVHN, already discussed in the main text. This skewed pattern does not arise in the **BP** models, whose representations contain fluctuations around their means, producing skewness coefficients much closer to 0. The value of  $-0.02$  for **BP** on MNIST (Layer 1) means that those activations are slightly negatively skewed.

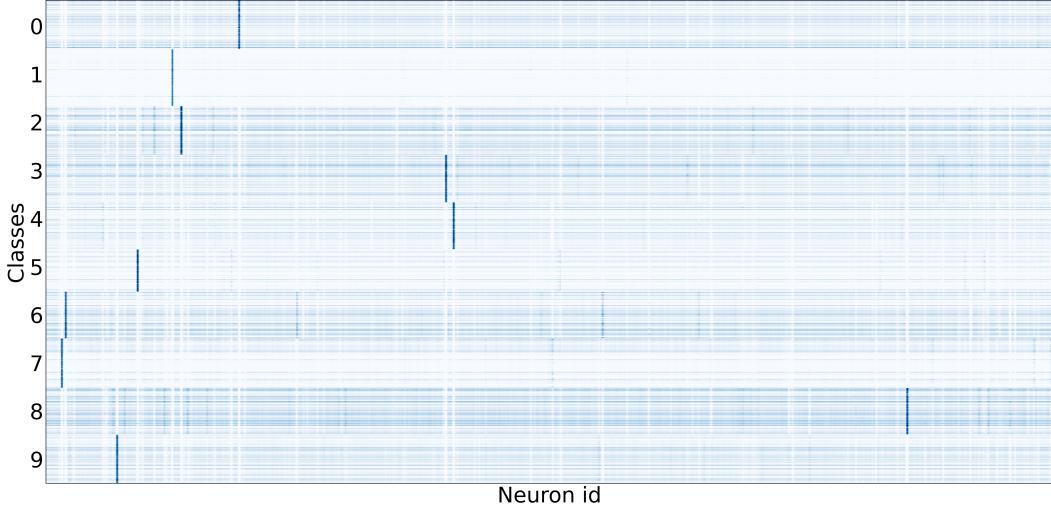


Figure 5: Activation patterns in a Multi-Layer Perceptron trained with the Forward-Forward algorithm, on the MNIST dataset. The image represents the activation map for neurons in Layer 2 for all images, grouped by class. A blue dot in position  $(x, y)$  indicates that neuron  $x$  is activated by input  $y$ ; colorscale represents the intensity of such activation (incorrectly classified samples have been removed). Horizontal bands mark different categories; dark blue vertical lines mark active neurons. Each input category activates consistently a specific sets of neurons (ensemble).

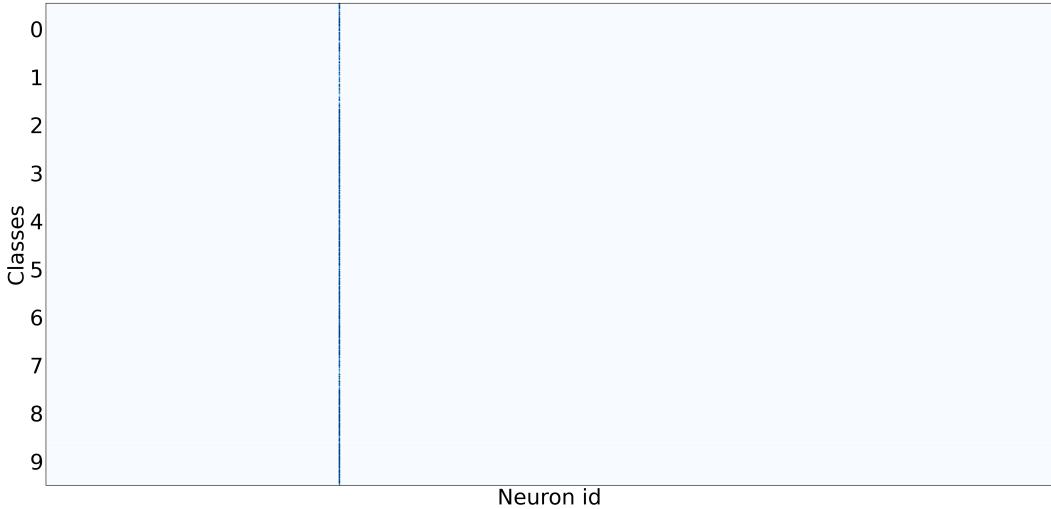


Figure 6: As in Figure 5, for Layer 3. Notice that there is only one unit that activates significantly and does not play a role in discriminating categories. The role of this layer, in this experiment seems, not related to the classification task.

## F Further results on representations of unseen categories and their ensembles

We showed in subsection 4.4 that a FF model trained on the FASHIONMNIST dataset – deprived of one category – can respond at test time to this unseen category with a valid ensemble (Figure 3). We report here the results of similar experiments, removing one category at a time. It turns out that, in six out of ten cases (we performed a single run for each category), the representations of the unseen category form a valid ensemble (Figure 7). It is with this situation in mind that we refer to “the ensembles related to unseen categories”.

Table 4: Average Fisher-Pearson moment coefficient of skewness of representations. Results expressed as  $\text{mean} \pm \text{std. dev.}$  over 10 runs with independent randomised weight initialisation.

Dataset	Layer	<b>FF</b>	<b>BP/FF</b>	<b>BP</b>
MNIST	1	$17.69 \pm 0.15$	$23.86 \pm 0.55$	$0.23 \pm 0.05$
	2	$6.89 \pm 0.52$	$4.71 \pm 1.21$	$0.09 \pm 0.03$
	3	$27.90 \pm 0.01$	$27.91 \pm 0.01$	$-0.02 \pm 0.02$
FASHIONMNIST	1	$18.66 \pm 0.25$	$26.85 \pm 0.82$	$0.38 \pm 0.06$
	2	$8.28 \pm 0.46$	$3.97 \pm 2.99$	$0.31 \pm 0.11$
	3	$24.09 \pm 0.94$	$27.40 \pm 0.79$	$0.05 \pm 0.06$
SVHN	1	$14.99 \pm 0.21$	$0.21 \pm 0.05$	$0.08 \pm 0.03$
	2	$5.82 \pm 0.28$	$15.15 \pm 0.43$	$1.90 \pm 0.42$
	3	$12.91 \pm 0.27$	$3.01 \pm 0.70$	$2.16 \pm 0.21$
	4	$4.95 \pm 2.94$	$11.79 \pm 0.98$	$0.85 \pm 0.02$

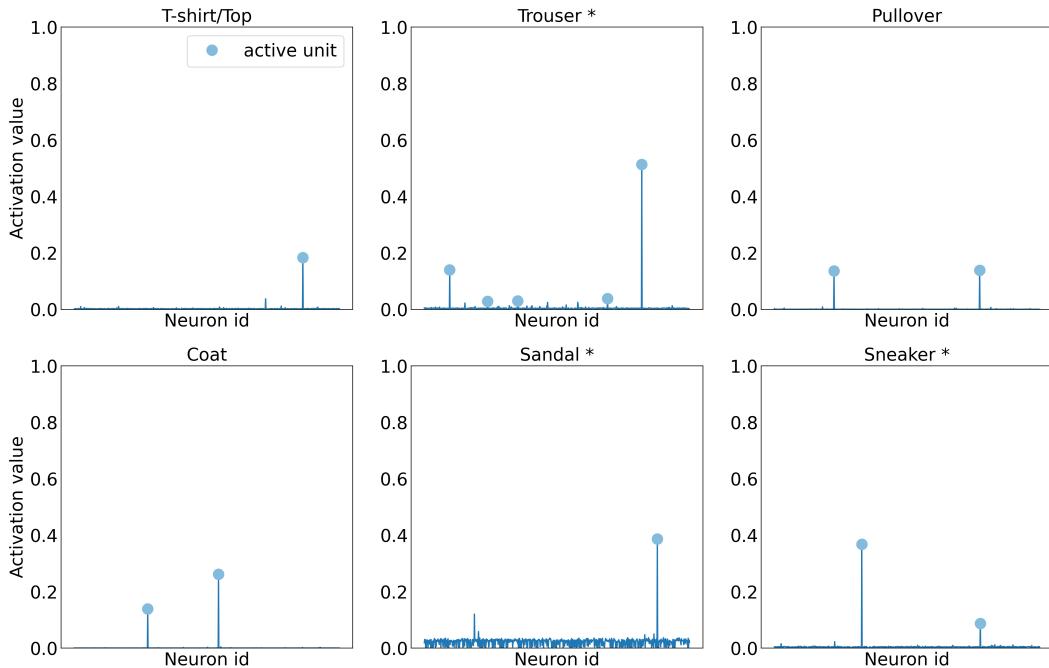


Figure 7: Ensembles elicited by the FF model trained on FASHIONMNIST deprived of one category. Activation value of each neuron in the first hidden layer (Layer 1), averaged on all images of the unseen category. Neuron index on the  $x$  axis; average activation on the  $y$  axis. Blue dots indicate units that are considered active according to the method described in subsection 3.5. Three out of ten elicit a strong ensemble and are marked with an asterisk (namely, Trouser, Sandal and Sneaker), other three elicit a less prominent but significant response (T-shirt/Top, Pullover and Coat). The remaining categories (Dress, Shirt, Bag, Ankle boot) do not elicit a significant response and are not reported.

When an unseen category forms a valid ensemble, it generally exhibits a high level of integration with the ensembles associated with the categories encountered during training. This integration implies that it can share common units with ensembles belonging to related categories. We show in Figure 8 how the ensembles of a selection of categories (T-shirt/Top, Pullover, Coat and Sandal) integrate – by sharing units – with the other ensembles.

The unseen categories in Figure 8 all share one characteristic: they all have a clear relationship (semantic/visual) with at least another category that has been used during training. On the contrary, the category Trouser is unique in the context of the FASHIONMNIST dataset, not being immediately

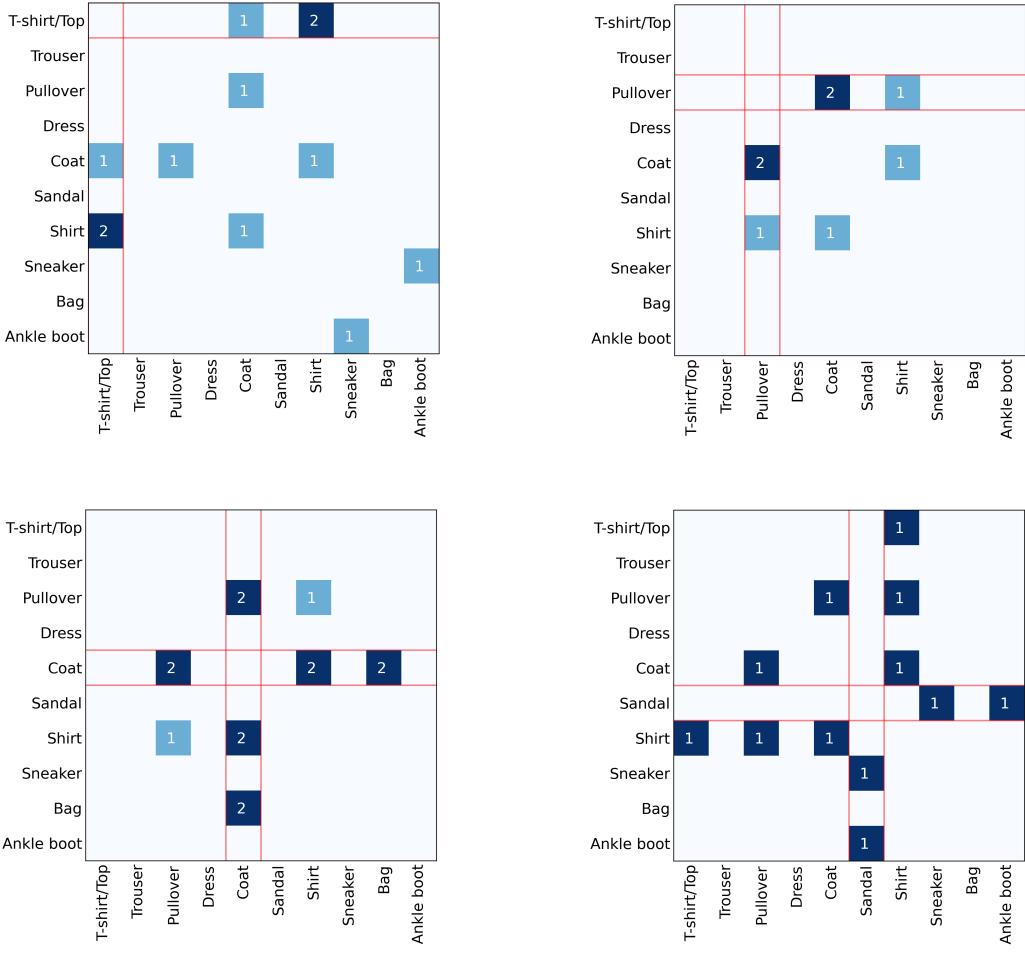


Figure 8: Shared units between the ensembles of unseen categories and the ensembles of categories seen during training (stripes delimited by the red lines). The results for T-shirt/Top, Pullover, Coat and Sandal are shown.

related to other categories by semantic/visual relationships. Nonetheless, when we train a FF deprived of this category we find at test time a strong ensemble signal (the most significant in this batch of experiments). We report in Figure 9 the characterisation of its activation pattern. This is particularly intriguing, because it shows that an unseen category can elicit an ensemble even in absence of a semantically/visually related category. Overall, these result relates to biological neural networks [16, 8], where ensembles appear to be the functional building block of brain representations even in the absence of known stimuli.

## G E/I imbalance for all models, layers and datasets

In subsection 4.5 we showed that all the models we considered give raise to very different networks, as it is shown in Figure 4, where we report marked differences, across all models, in the distribution of the ratio  $\varrho^+$  of positive weights. We report in Table 5 the average value of this ratio across all models, layers and datasets used. The data shown in Figure 4 refer to the second hidden layer of models trained on MNIST. We observe that in the first hidden layer all models are approximately close to balance, possibly due to the role of the first hidden layer in extracting low-level features from a normalised input.

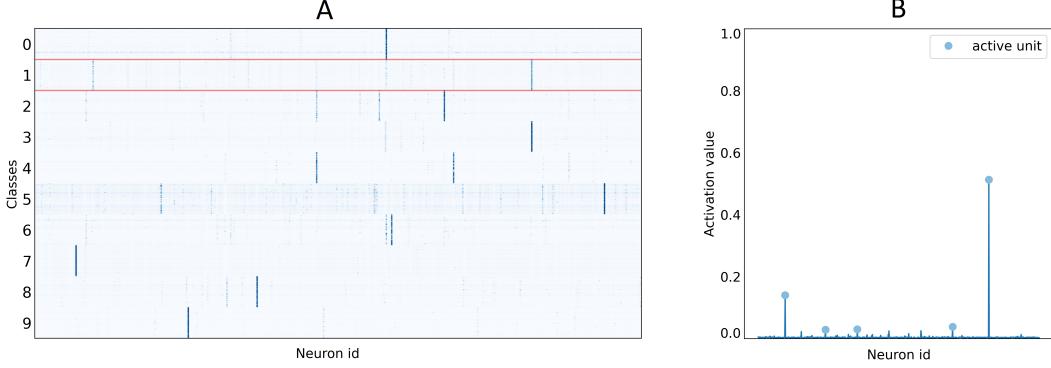


Figure 9: The representation of an unseen category can form a valid ensemble even in absence of a semantically/visually related category.

Panel A Activation pattern. The unseen category (`Trousers`) is surrounded by red lines.

Panel B Activation value of each neuron, averaged on all images of the unseen category.

Table 5: Fraction of positive weights ( $\varrho^+$ ) for each layer. Results expressed as *mean  $\pm$  std. dev.* over 10 runs with independent randomised weight initialisation.

Dataset	Layer	FF	BP/FF	BP
MNIST	1	$0.674 \pm 3.3 \cdot 10^{-3}$	$0.680 \pm 1.7 \cdot 10^{-2}$	$0.522 \pm 2.9 \cdot 10^{-3}$
	2	$0.039 \pm 1.8 \cdot 10^{-3}$	$0.454 \pm 3.1 \cdot 10^{-2}$	$0.504 \pm 1.6 \cdot 10^{-3}$
	3	$0.011 \pm 2.8 \cdot 10^{-3}$	$0.059 \pm 2.1 \cdot 10^{-2}$	$0.504 \pm 1.0 \cdot 10^{-3}$
FASHIONMNIST	1	$0.503 \pm 2.7 \cdot 10^{-3}$	$0.463 \pm 3.3 \cdot 10^{-2}$	$0.502 \pm 3.1 \cdot 10^{-3}$
	2	$0.039 \pm 1.9 \cdot 10^{-3}$	$0.478 \pm 4.7 \cdot 10^{-2}$	$0.493 \pm 3.9 \cdot 10^{-3}$
	3	$0.074 \pm 1.4 \cdot 10^{-2}$	$0.037 \pm 3.5 \cdot 10^{-2}$	$0.497 \pm 5.0 \cdot 10^{-3}$
SVHN	1	$0.492 \pm 1.2 \cdot 10^{-3}$	$0.499 \pm 1.6 \cdot 10^{-3}$	$0.500 \pm 1.2 \cdot 10^{-3}$
	2	$0.058 \pm 2.1 \cdot 10^{-3}$	$0.270 \pm 2.4 \cdot 10^{-2}$	$0.468 \pm 4.5 \cdot 10^{-3}$
	3	$0.269 \pm 3.3 \cdot 10^{-2}$	$0.387 \pm 4.2 \cdot 10^{-2}$	$0.414 \pm 1.0 \cdot 10^{-2}$
	4	$0.131 \pm 7.8 \cdot 10^{-2}$	$0.019 \pm 3.0 \cdot 10^{-2}$	$0.429 \pm 6.7 \cdot 10^{-3}$