Reviewer 1: We thank the reviewer for a thorough review and kindly request a reevaluation considering the following. *Relation to Xie and Seung's contrastive Hebbian learning (CHL):* Indeed a missed opportunity! Thank you for bringing this up. We will provide a detailed comparison. However, we think the relation is not as direct as the reviewer suggests. In deep similarity matching, error is defined as a function of all neurons at all layers, and through duality can be reduced to a local error for each synapse. In this sense, there is nothing special about feedback connections. Even without the feedback ($\gamma = 0$ limit), each layer is doing gradient-based learning. This is in contrast to CHL, which optimizes an error defined at the output, and therefore has to (back)propagate it by feedback connections. Some other differences are: 1) CHL performs approximate gradient-descent. Our network performs exact gradient descent-ascent. 2) Our network has lateral connectivity, CHL does not. 3) CHL has clamped and unclamped phases, our network does not.

Similarity matching as a deep learning objective function: This is a very fair criticism and we should have been very clear about it in our paper. The name similarity matching suggests that pairwise similarities are preserved across layers, but in reality, because of the structured connectivity and regularization, the similarity structure changes in a very nonlinear way. For example, previous work showed that (unstructured) nonnegative similarity matching can be interpreted as nonnegative ICA or manifold learning. By introducing structure and depth, surely the network's power to warp similarity metrics will increase. An analytical understanding of how exactly such warping will happen is hard, and should be the topic of later papers. Our simulations were chosen to give some intuition: one illustrates how feedback can provide associative links, and the other one shows hierarchical feature extraction. We thank the reviewer for the autoencoder connection, which we will make.

Comparison to Bahroun et. al. While we appreciate Bahroun *et. al.*'s important contribution, our paper's scope goes much beyond it. To be technically correct, Bahroun *et. al.*'s network uses (biologically-implausible) weight sharing, and therefore it is basically a repeated set of (unstructured) similarity matching networks tiling an image. This is different than our locally structured network, which can learn different features for different portions of an image, while possibly having long-range lateral interactions. When citing Bahroun *et. al.* we did not make this difference very clear, and probably caused the reviewer's confusion. We apologize for that. More importantly, we provide global objective functions for a much larger family of structured and deep architectures (with local learning), locally-connected being just one example.

(NpS, d_l)	(4,0)	(4,4)	(16,0)	(32, 0)
error (%)	4.96	4.59	2.79	2.05

γ	0	0 (n)	0.025	0.025 (n)
error (%)	8.12	9.53	7.67	8.85

Table 1: MNIST classification (test error) by a linear classifier trained of representations learned by a locally connected network. Top: single-layer, bottom: two-layers. Neurons are organized with a stride of 2 and multiple neurons (or features) per site (NpS), got feedforward input from pixels within radius 4, and lateral input from neurons within radius $\leq d_l$. Two-layer simulations had NpS = 2 and $d_l = 4$ for the first layer and 8 for the second. (n) denotes test set (not training) with occluded 3x3 patches. Increasing NpS and d_l increased the performance of the classifier. Optimal γ is > 0.

Empirical results: All reviewers correctly asked us to asses the quality of our learned features by a classification task. We launched a detailed numerical study, some preliminary results on MNIST is in Table 1. CIFAR-10 will be added. *Weight transport:* We appreciate the concern about our solution's robustness. We haven't done a detailed robustness analysis (except to initialization), and we will discuss this point. We recently learned about similar ideas in the backpropagation literature, which we will cite (Kevin and Pollack 1994, Akrout *et. al.* 2019).

Reviewer 2: We thank the reviewer for the enthusiastic support! In Eq. (2) y_i is the same variable as r_i. Sorry!

Illustrative example: We will provide details here, thanks! To be fair, all relevant information about the dataset for similarity matching purposes is already in the shown similarity matrices in the figure (input data have dot products of $\sim$0.5 across clusters, and $\sim$1 within the cluster), but we should have made our data generation clear and will do so.

Figure 3: We used a common technique from neuroscience for visualization: reverse correlation, which is an average of the input images weighted by a neuron's response. We then set to zero the portions of the image that will not elicit a response in the neuron, because of the limited range of connectivity. We will expand on this in an appendix. Interpretation: Closest would be ICA due to nonlinearity, please also see response to Reviewer 1.

Classification results: Please see our response to Reviewer 1 and preliminary results in Table 1.

Global lateral inhibition: This is an excellent suggestion! With fixed weights, the network ends up solving a modified similarity matching problem, Eq. (8) with L fixed. It is not clear to us whether one approach is better than the other. We will cite the Krotov, Hopfield reference (which we should have already done) and discuss this point.

Reviewer 3: We thank the reviewer for the sincere feedback, which was truly a wake up call. We were embarrassingly reminded that our mathematical notation and terminology that was lucid to us because of our scientific background may not be so for others. We will revise our paper with this in mind, and take into account all your suggestions. In particular, section 2 was meant to be a short review of previous papers augmented with new findings. This strategy does not work. We will expand section 2, and others, to make the paper self-contained, moving technical details to the appendix.

Hyperparameters and empirical results: Please see our response to Reviewer 1 and preliminary results in Table 1.

Originality: The high level ideas of depth and structured connectivity have existed much before they appeared in the sparse coding literature. Our contributions are 1) implementing these ideas in the similarity matching framework at the cost function level, and 2) providing biologically-plausible networks that can optimize such cost functions.