MODELING VISUAL INVARIANCE WITH SYMMETRY REGULARIZATION

Aishwarya H. Balwani
Graduate Student, ECE

Lucas Kiefer
Undergraduate Student, CS

Somnath Sarkar
Graduate Student, CS

Introduction

While computers have made drastic leaps in recent years at tasks such as object recognition [1], humans – and other "natural" intelligent agents – are still able to recognize objects across translations, rotations, noise, and obfuscations far more *consistently* than their artificial counterparts. Consequently, being able to reach and beat human performance reliably when recognizing objects in the presence of nuisance variables or factors that ought to count as *invariances* as far as the recognition task goes, has been and continues to be an area of active research.

This project looks into one aspect of this rather broad problem, and proposes to learn representations of 2D shapes such that they are *equivariant* to any transformation that can be represented by the actions of an arbitrary finite *group*. Towards this end, we encourage a group structure on the weights of a dictionary being used to represent the stimuli using a regularizer that enforces the learnt solution to respect the symmetries in the unlabelled training data. In particular, we look to induce equivariance in representations when our stimuli are subjected to cyclic rotations. Our experiments provide proof of concept that it is indeed possible to induce a certain degree of equivariance in the learnt representations (at least when the transformations can be represented by simple, finite groups) without explicit knowledge of the underlying symmetries in the data. Finally, we also discuss the biological plausibility of such a learning scheme.

Human Proficiency

While fairly simplistic versions of tasks like these had been difficult for computers to master until the rise of convolutional neural networks [2, 3], human and mammalian brains are able to do well on such tasks with relative ease. Humans consistently exhibit transformation *invariance* (notably a special case of equivariance), when recognizing shapes as the same regardless of scale, rotation and translation. This has been shown in multiple different studies, with humans learning this skill at a very young age ($\sim$ 3-4 months) [4] and continuing to maintain proficiency on the task through most of their lifetimes. Moreover, in addition to humans, other animals have also shown a clear proficiency in this context. For example, Zoccolan [5] showed that rats too are capable of invariant visual object recognition, being able to correctly and consistently identify the same object as the identical, regardless of translation and rotation.

Figure 1: *Modelled stimulus class.* Examples of the three different shapes (square, ellipse, heart) and their corresponding translations which are encouraged to have equivariant representations via symmetry regularization.

Modeling the Stimulus Class

Our stimuli are 16×16 binary images with black being 0 and white being 1. They are generated by first choosing a template 64×64 binary image for each of the three base shapes in the dSprites dataset [6] and downsampling the same by a factor of 4 along each axis. We then produce all possible cyclically translated versions of an image I as follows:

1. Construct the down-shift matrix $D \in \{0,1\}^{m\times m}$ where $D_{ij} = \delta_{i,(j+d)\%m}$
2. Similarly[1] construct the right-shift matrix $R \in \{0,1\}^{n\times n},$ where $R_{ij} = \delta_{(i+r)\%n,j}$
3. $I_{rd} = DIR.$

Since both $r, d \in \{0, 1, ...15\}$ independently, we have 256 cyclically shifted images of each shape. Combining all cyclically translated versions of the three base shapes, our dataset has a total of 768 images (Figure 1).

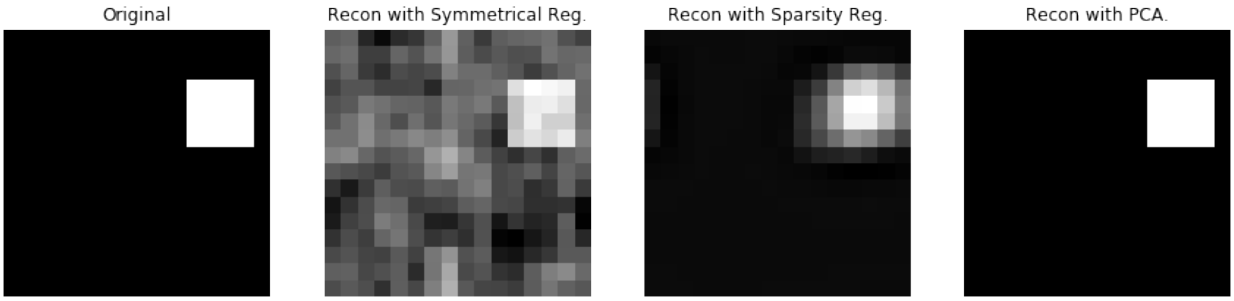

Figure 2: *Qualitative reconstruction results.* An example of i) An original data sample, and its reconstruction using a dictionary learnt with ii) the group-theoretic symmetry regularizer, iii) sparse coding, and iv) PCA.

Learning Algorithm

The matrices R, D and their products gives us a set $\mathcal{G}$ with $|\mathcal{G}| = 16 \times 16 = 256$ since the number of unique cyclical translations possible (which by definition would be the cardinality of $\mathcal{G}$) is mn when the input image they act on is of the shape $m \times n$. Moreover, we note that $(\mathcal{G}, *)$[2] is a *group*. This group structure is due to the fact that i) The individual matrices R, D and their subsequent products are orthonormal matrices that form a closed set under matrix multiplication, and ii) The identity matrix I_{mn} is also a member of the set, corresponding to zero pixel shifts.

The learning algorithm we use [7] consequently looks to directly learn equivariant representations of data that are acted upon by elements of the same finite group $\mathcal{G}$ with the help of regularizations that leverage data symmetry. In particular, for a general learning algorithm that minimizes a loss function of the form

$$\min_{W} \mathcal{L}(W, S_N) + \beta\mathcal{J}(W) \tag{1}$$

where W are the atoms of a dictionary, S_N are the training data, $\mathcal{L}$ is a generic loss function, and $\beta \in \mathbb{R}_+$ is a free parameter, the regularizer $\mathcal{J}$ is used to restrict the weights in W to be symmetry adapted. We assume that our input data $S_N = \{x_i\}_{i=1}^{N} \subset \mathcal{X}$ is a finite collection of exactly Q complete orbits w.r.t. a single group $\mathcal{G}$, thus implying that $N = |G|Q$. This in turn requires that our dictionary $W \in \mathbb{R}^{d\times|G|}$ with d being the dimension of our inputs.

Further elaborating on our regularizer $\mathcal{J}$, our complete objective function is formulated as

$$\min_{W} \mathcal{L}(W, S_N) + \gamma||[XX^T, WW^T]||_F^2 + \beta r(W^TW) \tag{2}$$

with $\mathcal{L}$ being our loss for representation learning, i.e., reconstruction error as measured by pixel-wise mean square differences. The second term[3] encourages the Gramians of the training data and dictionary atoms to commute, thus ensuring that the weights W are of the same latent group as that in the training set distribution. Finally, the function r calculates all the differences between the components of the vectorized form of $W^TW \in \mathbb{R}^{|\mathcal{G}|^2}$, thereby promoting equivariance in the atoms of W. Both $\gamma, \beta \in \mathbb{R}_+$ are free parameters.

Simulations and Results

We conducted simulations on three different types of data – i) A synthetic dataset that sampled data in $\mathbb{R}^6$ belonging to the cyclic group C_6, ii) A single orbit dataset consisting only of all the cyclically translated squares, and iii) A multi-orbit dataset consisting of data from cyclically translated squares, ellipses, and hearts.

[1] For both matrices R and D, δ_{ij} is the Kronecker delta and $\%$ represents modulo.

[2] Here $*$ denotes matrix multiplication.

[3] $[A, B] = AB - BA$ is the commutator operation between two square matrices of the same dimension.

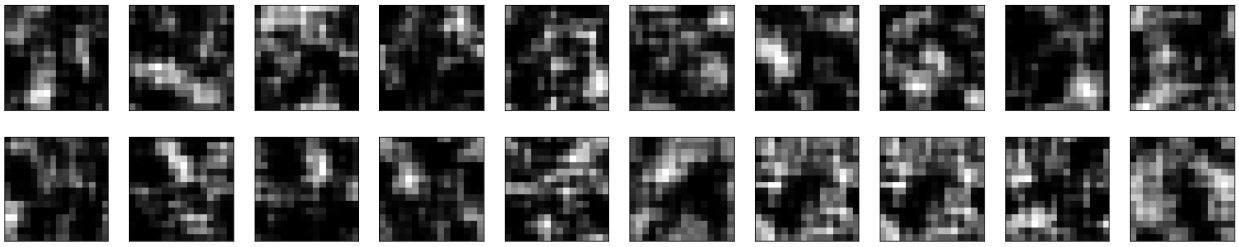

Figure 3: *Equivariance inducing dictionary atoms.* Randomly chosen examples of the atoms learnt using the symmetry regularizer with the translated squares data.

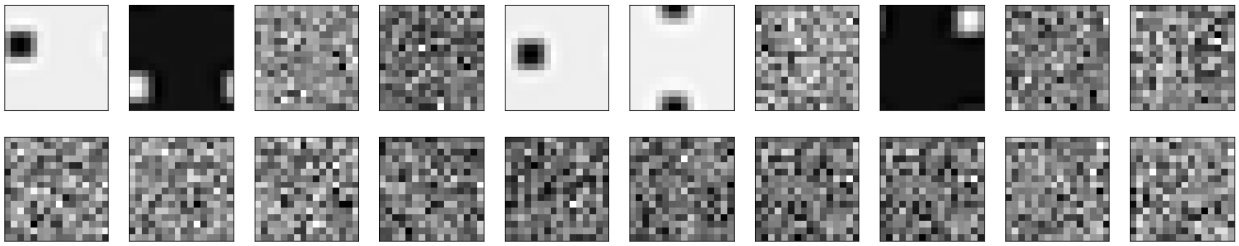

Figure 4: *Sparsity inducing dictionary atoms.* Randomly chosen examples of the atoms learnt using sparse coding with the translated squares data.

We learnt representations on all three datasets using three different objectives, viz., i) The one with group-theoretic, symmetry regularization, ii) Sparse coding, i.e., sparsity-based regularization, and iii) No regularization, i.e., where the only term in the objective function is reconstruction. This was done using PCA.

We then compared these three representation learning schemes on how well they reconstructed that data and also displayed representational equivariance, by measuring a permutation invariant pseudometric given by Euclidean distance between the histograms of all the translated versions of the input.

Equivariance Divergence	**Symmetry**	Sparsity	PCA
Cyclic Group C_6	**0.180**	0.182	1.109
Translated Squares	**0.744**	0.966	1.296
All Shapes	**0.220**	0.305	1.156

Reconstruction Error	**Symmetry**	Sparsity
Cyclic Group C_6	**0.137**	0.147
Translated Squares	0.240	**0.136**
All Shapes	0.146	**0.080**

Table 1: Comparison of divergence from equivariance in the learnt representations (left), and MSE of the reconstructed data (right) across different datasets and representation learning techniques.

In terms of reconstruction, as expected PCA with a dictionary having the same number of atoms as the data dimension does perfectly (Figure 2). However, PCA did the worst in terms of promoting equivariance in the learnt representations (Table 1). Figure 5 shows the top 20 principal components, which act as the atoms in the case of PCA.

The group-theoretic regularizer did comparably to sparse coding when reconstructing data from the group C_6 but didn't do as well when doing the same with real data. It should be noted though that the gap in performance dropped when more (i.e., multi-orbit) data were added thus making it difficult for both learning mechanisms to respect the constraints of their regularizers and the data reconstruction simultaneously. In terms of promoting representational equivariance, the group-theoretic regularizer does the best across the board, albeit by a small margin at times (Table 1).

A few randomly chosen atoms for both, the symmetry regularized and sparse coding schemes are shown in Figures 3, 4 respectively. It is interesting to see that for the case of sparse coding, there are a few elements which look very much like the data template themselves, but the rest seem like Gaussian noise. With the symmetry regularization on the other hand, most atoms look like curve detecting filters of some sort.

Biological Plausibility

As evidenced by our results, this method holds promise in being able to produce equivariant representations for transformed versions of an input that belong to the same group, thus getting us a step closer to the visual invariance

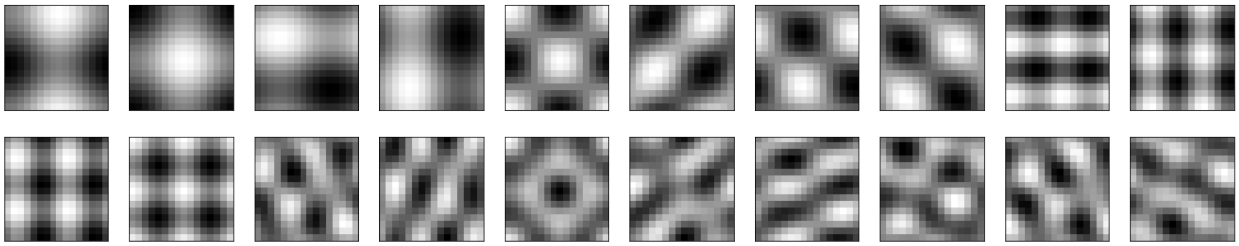

Figure 5: *Unbiased dictionary atoms.* Top 20 principal components learnt using PCA with the translated squares data.

capabilities of humans. Given that there has been support for sparse coding being implemented by the brain [8], we believe that the above described learning mechanism could indeed biologically plausible for learning group-equivariant representations since none of our regularizers affect the general sparse coding scheme. Finally, our learning is gradient-based, making use of the popular iterative shrinkage and thresholding algorithm (ISTA) [9], and while there has been some discussion on if and how the brain implements gradient descent or similar error-based learning mechanisms [10], we believe that it is definitely a possibility.

However, the above mechanism leaves much to be desired in terms of robustness in the presence of noise or absence of data, i.e., when we have access only to subsets of an orbit. Both intrapolation and extrapolation are skills that humans are very good at, and imbibing the same in this particular learning scheme would be important. The current mechanism also fails if data from multiple groups are presented together, and would need to be addressed for the sake of both, robustness and efficient learning.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[4] Mayu Nishimura, Suzy Scherf, and Marlene Behrmann. Development of object recognition in humans. *F1000 biology reports*, 1, 2009.

[5] Davide Zoccolan. Invariant visual object recognition and shape processing in rats. *Behavioural brain research*, 285:10–33, 2015.

[6] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[7] Fabio Anselmi, Georgios Evangelopoulos, Lorenzo Rosasco, and Tomaso Poggio. Symmetry regularization. Technical report, Center for Brains, Minds and Machines (CBMM), 2017.

[8] Michael Beyeler, Emily L. Rounds, Kristofor D. Carlson, Nikil Dutt, and Jeffrey L. Krichmar. Neural correlates of sparse coding and dimensionality reduction. *PLOS Computational Biology*, 15(6):1–33, 06 2019.

[9] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[10] Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, 10:94, 2016.