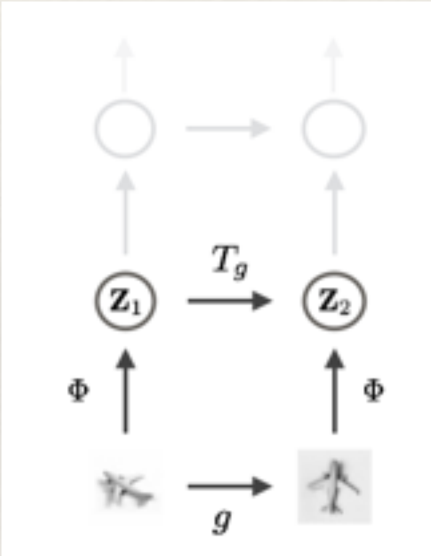
Group Equivariant Convolutional Networks

Taco Cohen

Outline

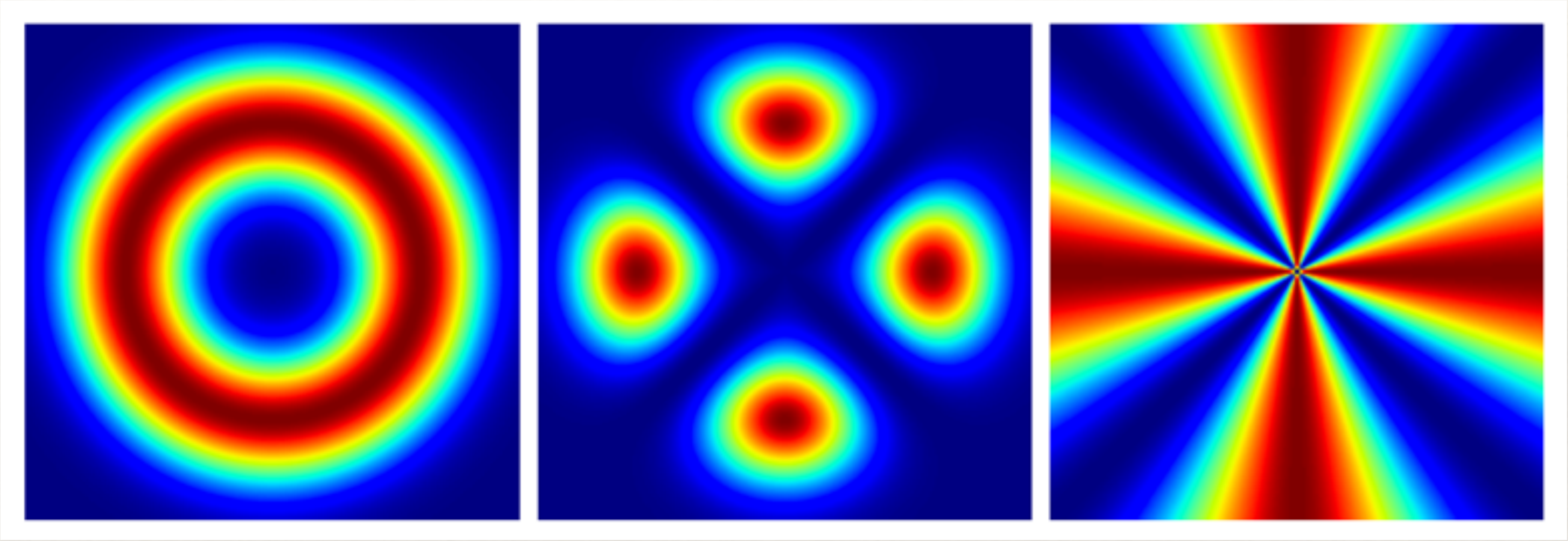
1. **Symmetry & Deep Learning:**
    - Statistical Power from Symmetry
    - Invariance vs Equivariance
    - Equivariance in Deep Learning
2. **Group Theory**
    - Symmetry, Groups
    - Subgroups, Cosets, Quotients
    - Wallpaper groups
3. **Group Equivariant Networks**
    - CNNs and translation equivariance
    - G-Convolutions
    - Equivariance of non-linearities
    - Equivariance of G-pooling operator
    - Backpropagation
4. **Algorithms**
    - Spatial & Spectral G-convs
5. **Results**



Background:
Invariance, Equivariance & Symmetry

Symmetry in ML

A symmetry of a function is a transformation that leaves that function invariant

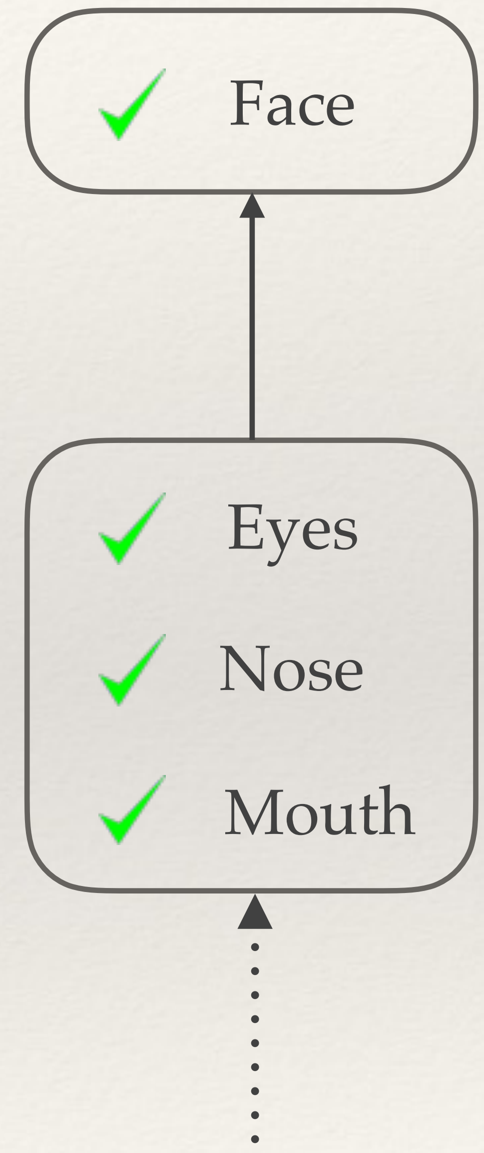


In ML: look for symmetries of densities, factors, label functions, …

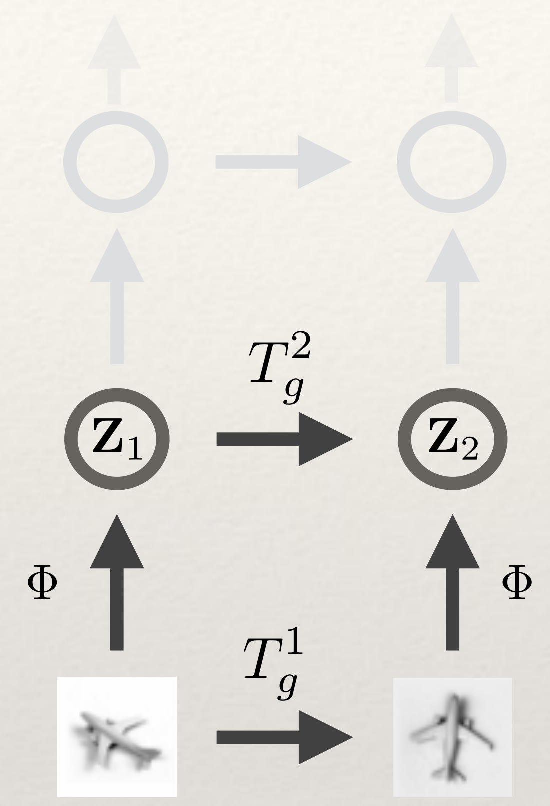
Shariff, R. (2015). Exploiting Symmetries To Construct Efficient Mcmc Algorithms.

Invariance

The “Picasso Problem”



Equivariance



$$\Phi(T_g^1\, x) = T_g^2\, \Phi(x)$$

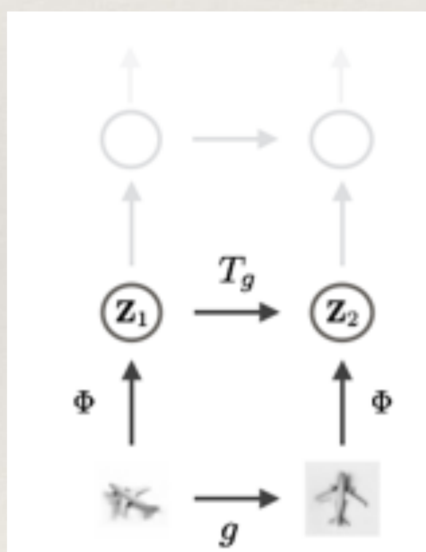
Hinton, G., Krizhevsky, A., & Wang, S. (2011). Transforming auto-encoders. ICANN-11

Lenc, K., & Vedaldi, A. (2015). Understanding image representations by measuring their equivariance and equivalence (CVPR)

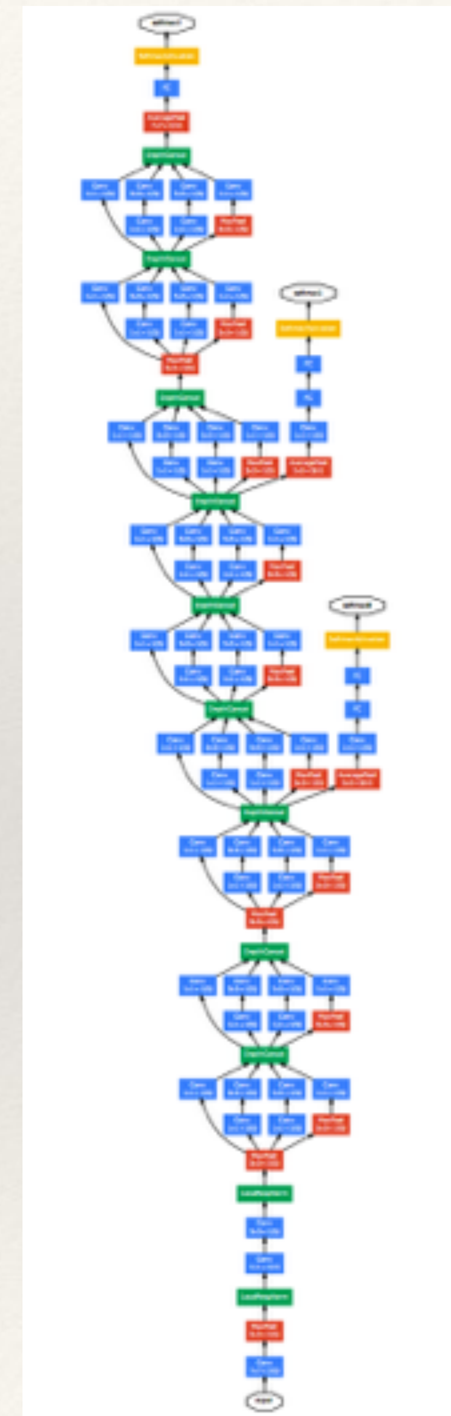
Cohen, T., & Welling, M. (2014). Learning the Irreducible Representations of Commutative Lie Groups. (ICML)

Symmetry in DL

- Why do CNNs work so well?
- They exploit *translational symmetry*
- In deep nets, each layer should preserve the symmetry



- The representation Φ should be an *equivariant map* for the symmetry group.



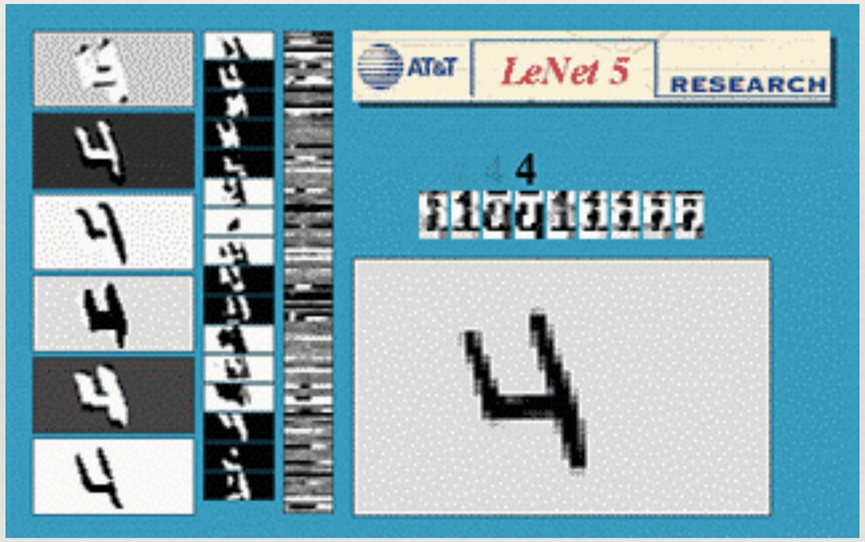
Deeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeep!

Szegedy et al. (2015)

ConvNets are Translation Equivariant

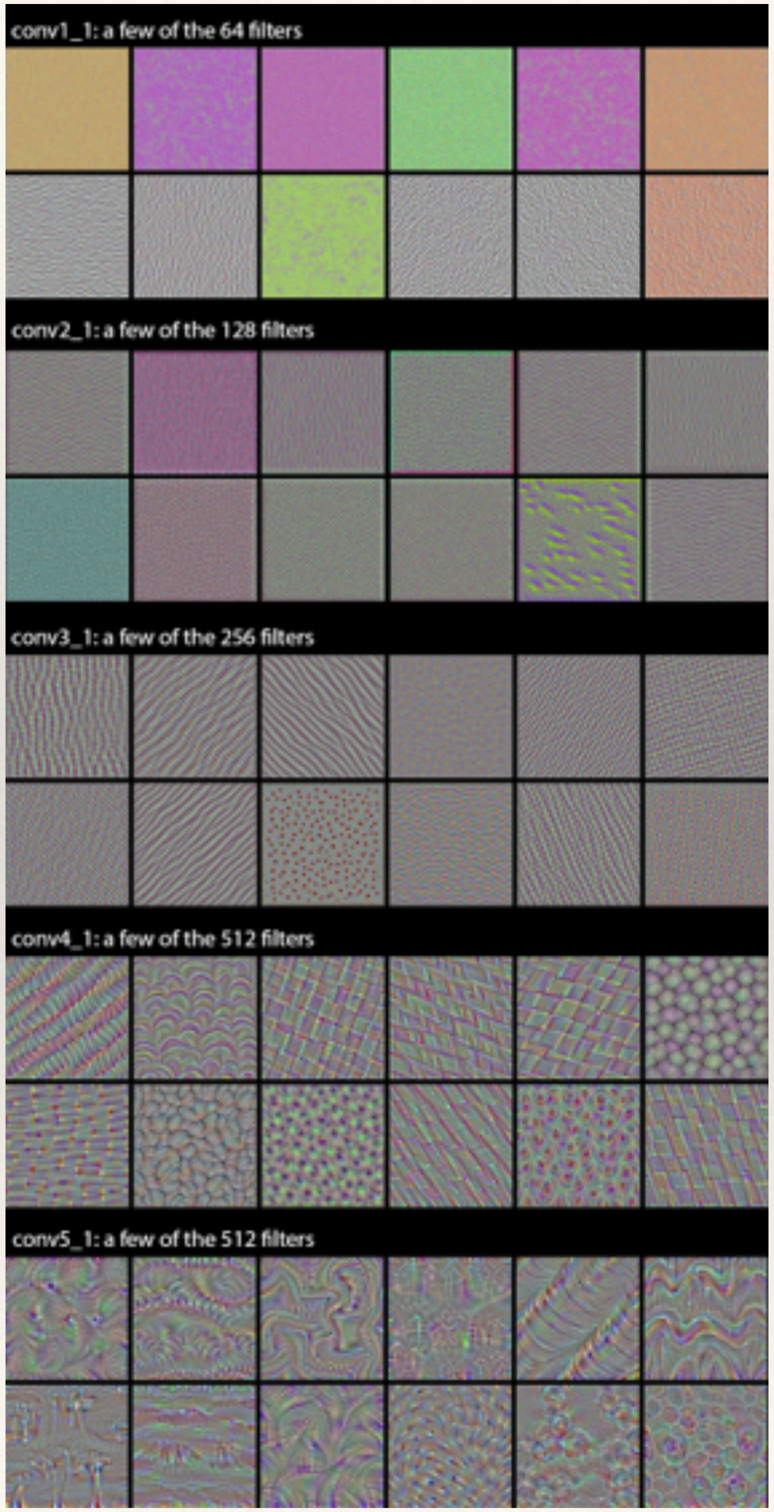
Source: http://yann.lecun.com/exdb/lenet/index.html

Are ConvNets Rotation-Equivariant?



Source: http://yann.lecun.com/exdb/lenet/index.html

CNNs want to be Equivariant

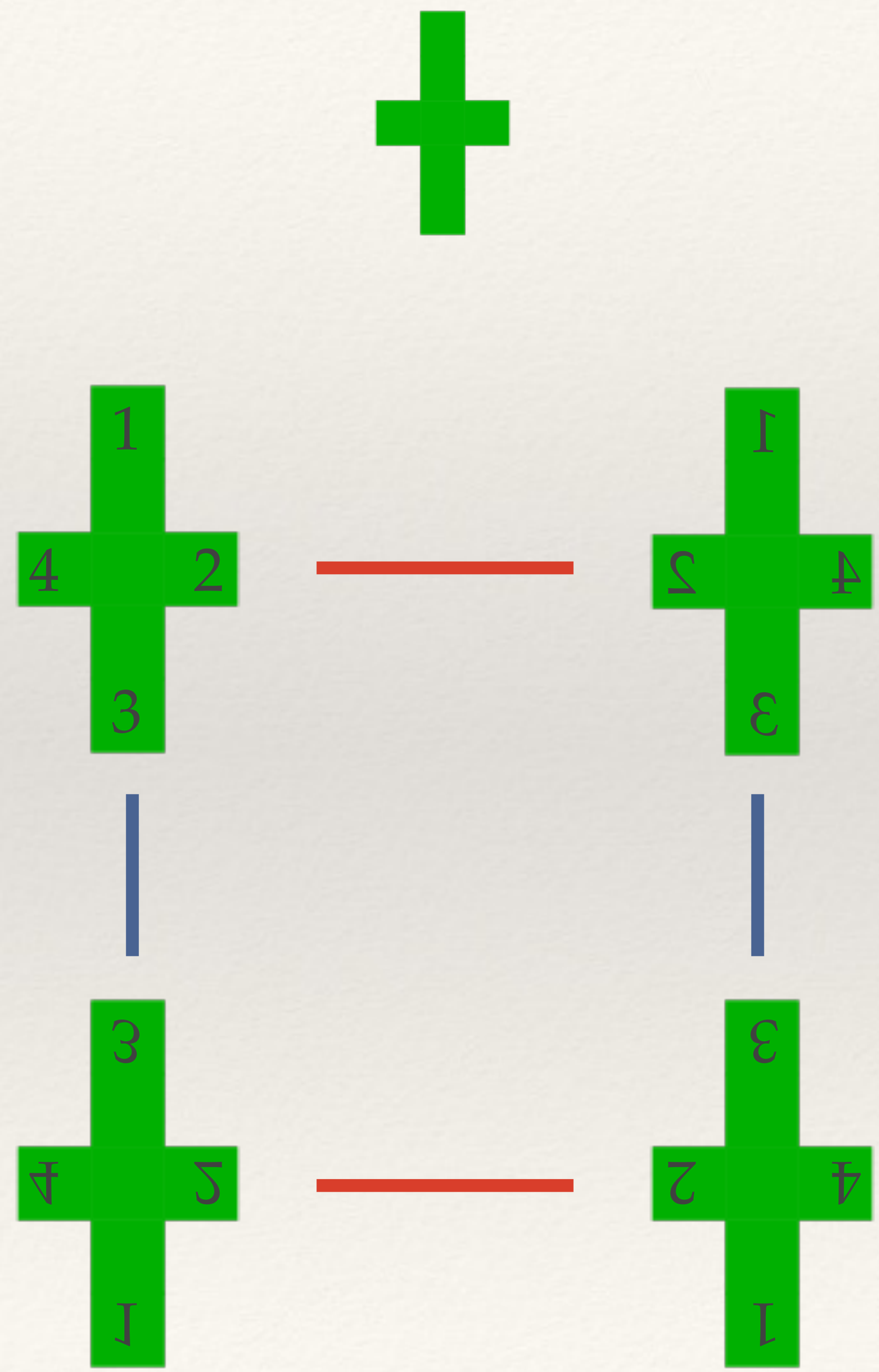


http://blog.keras.io/how-convolutional-neural-networks-see-the-world.html

Visual Group Theory

With figures from “Visual Group Theory” by Nathan Carter (2009)

Symmetries



Groups

A group is:

- a set, G,
- together with a binary operation that combines two elements a, b in G and produces another element ab,
- that satisfies the group axioms:

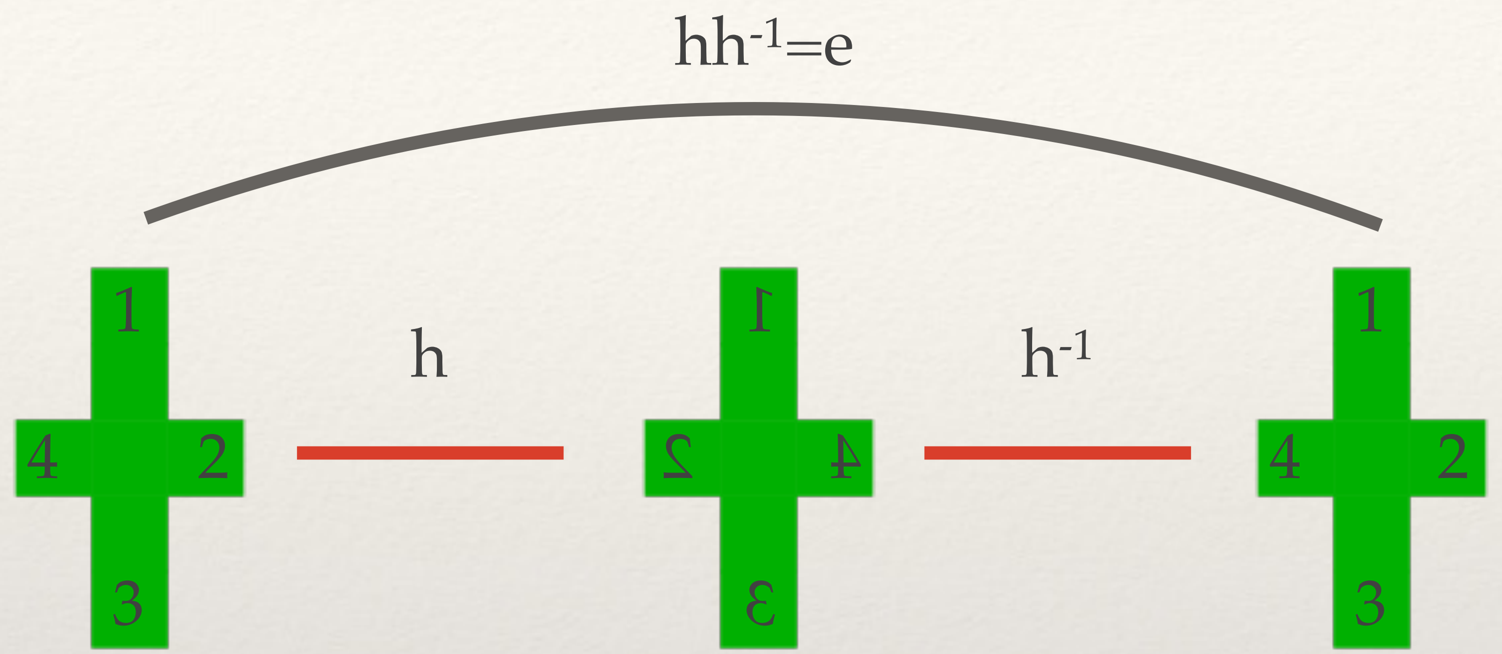
1. **Identity:** there exists an element e in G, such that for every element a in G, ea = ae = a

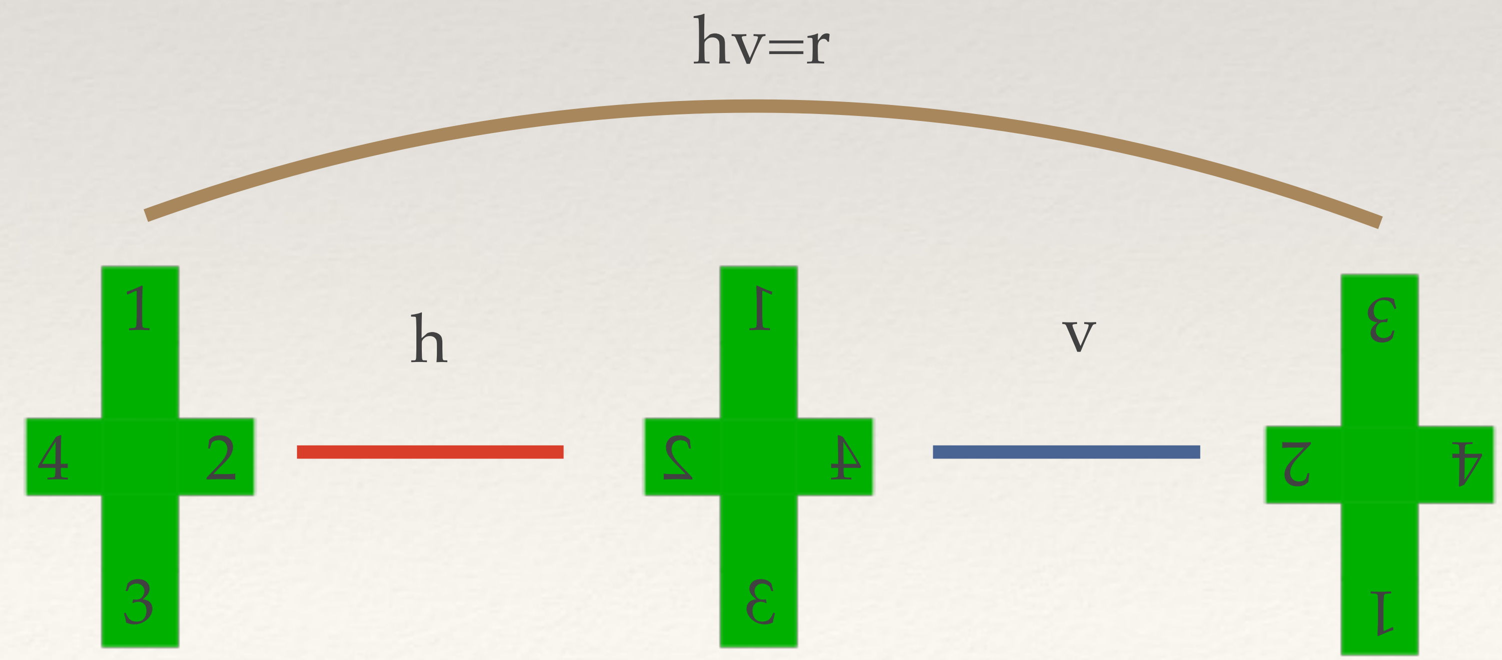
2. **Associativity:** For all a, b and c in G, (a b) c = a (b c).

3. **Closure:** for all a, b in G, the composition ab is also in G

4. **Inverse:** for each a in G, there exists an element b in G such that ab = ba = e

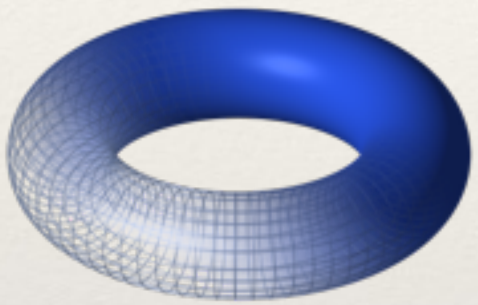
The Symmetries of an Object form a Group



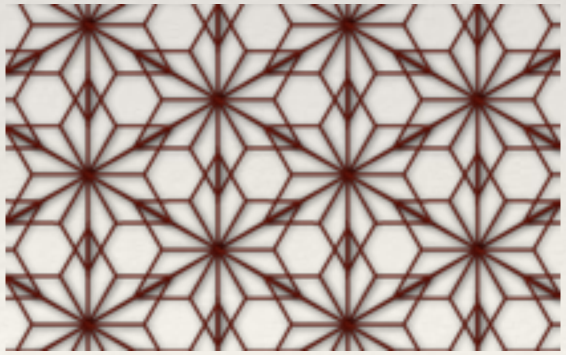


Examples

Lie Groups

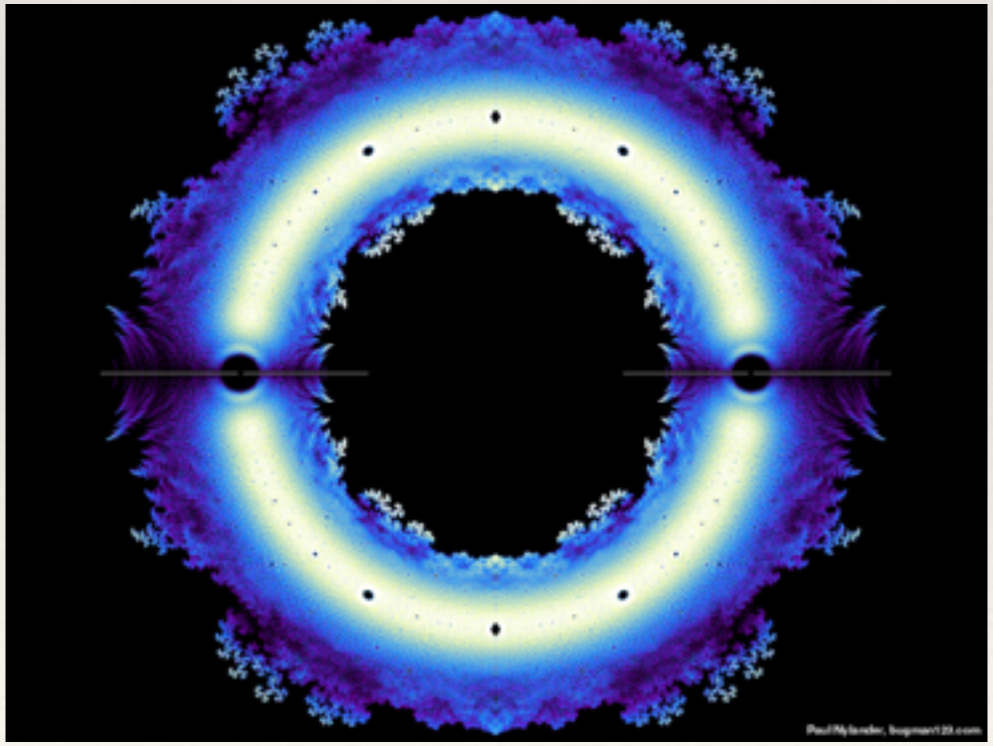


Wallpaper Groups

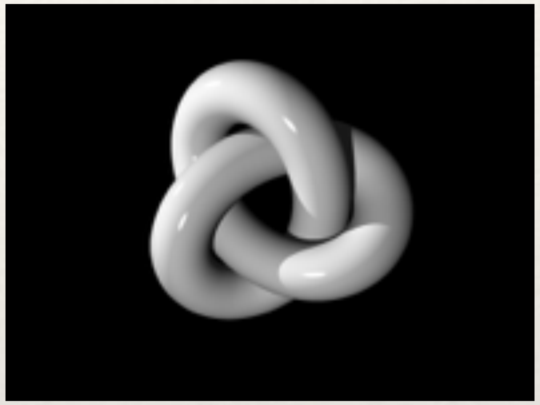


Discrete Groups

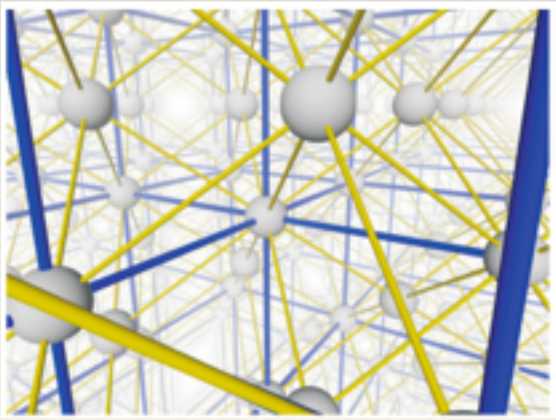
Galois Groups



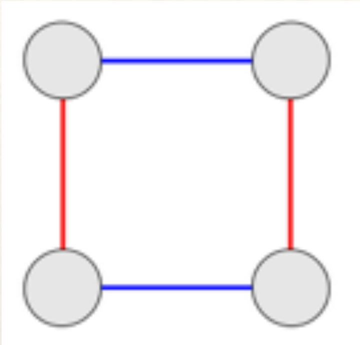
Topological Groups

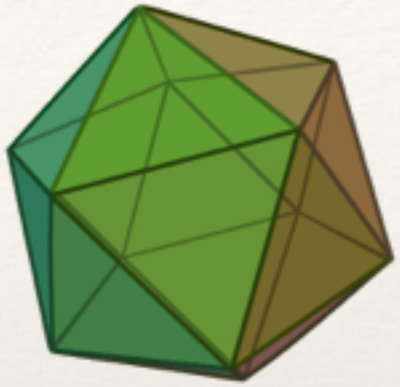


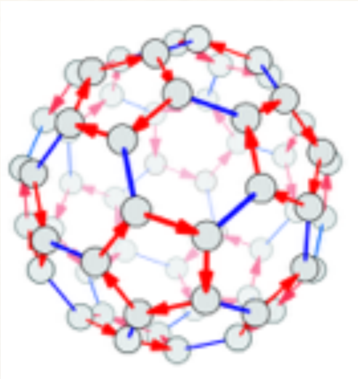
Space Groups

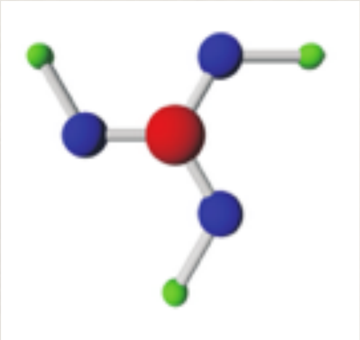


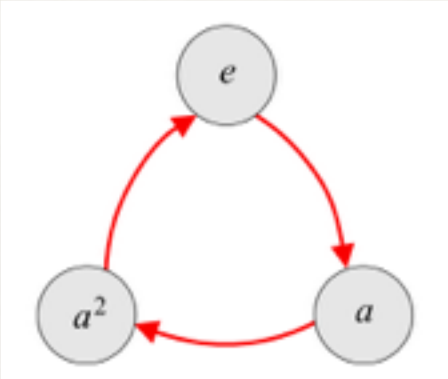
Cayley Graphs

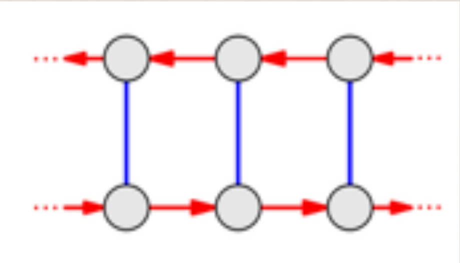


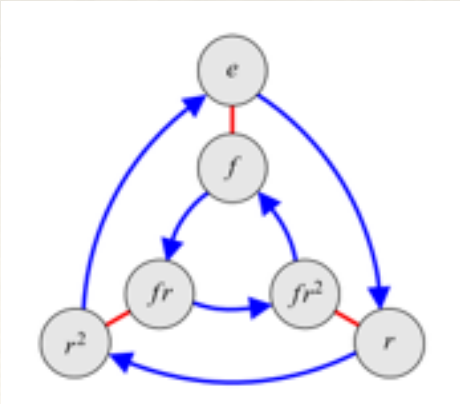


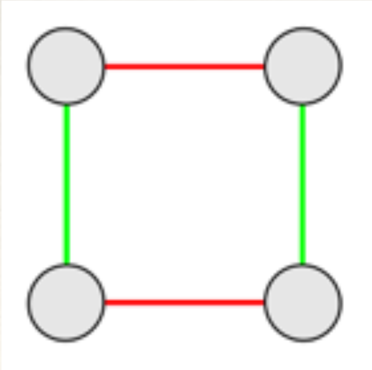


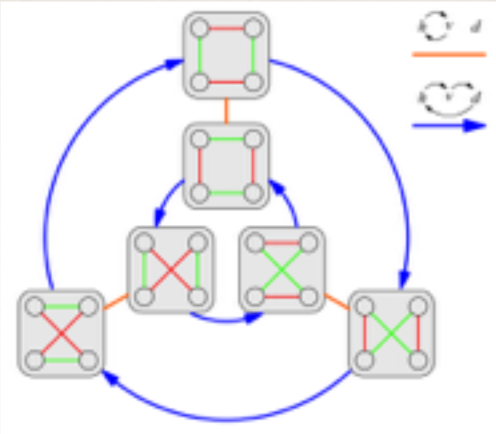






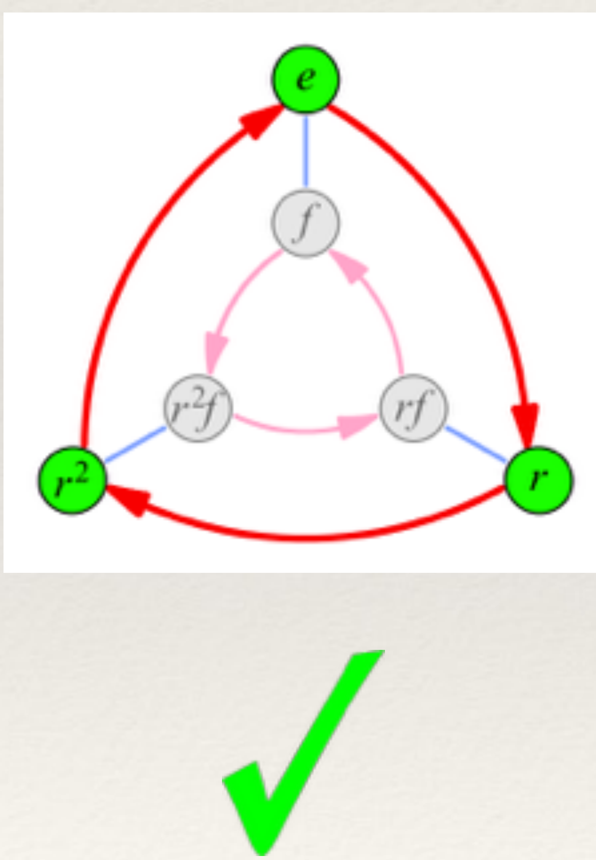


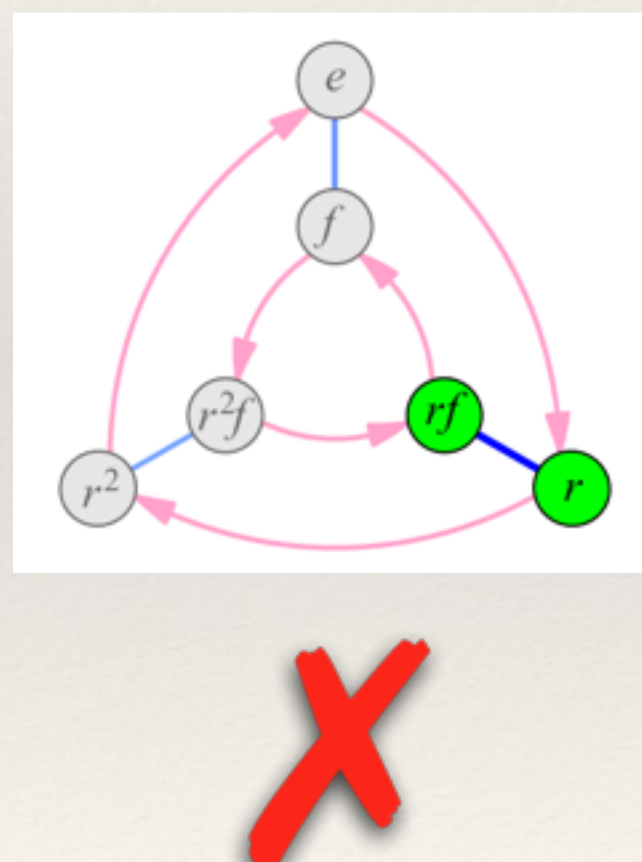


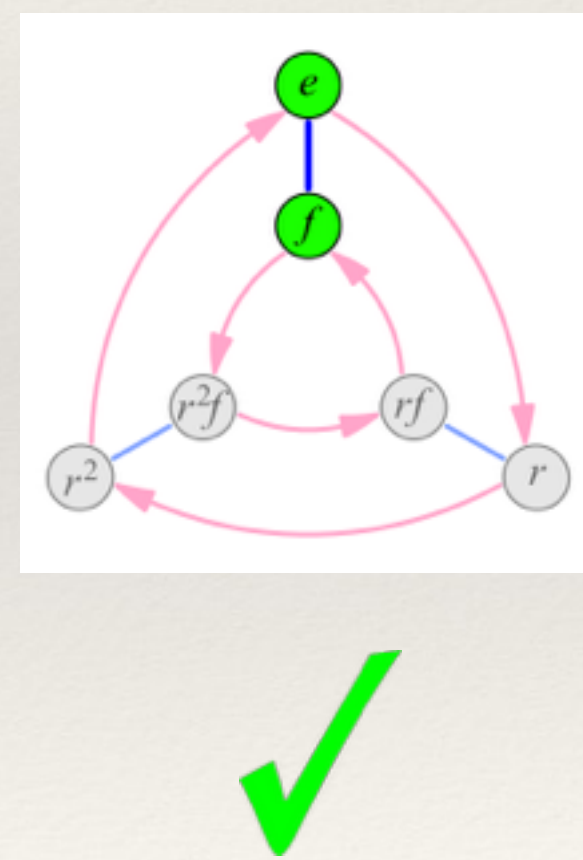


Subgroups

A *subgroup* of a group is a *subset* of said group, that is itself a group



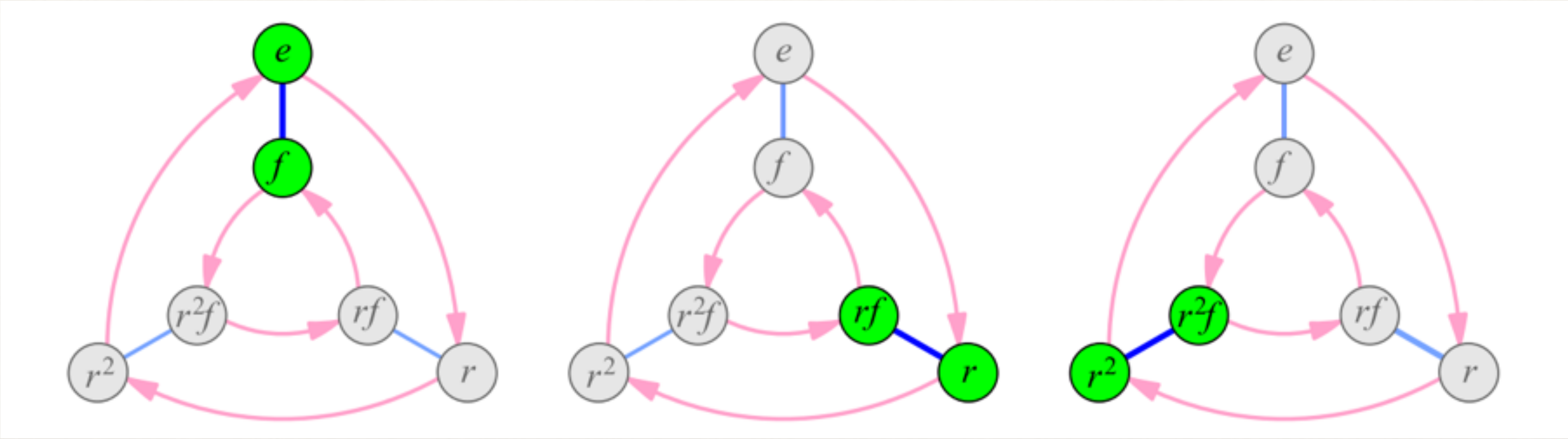




Cosets

The left coset of subgroup H in G with respect to g is the set:

$$gH = \{gh \,|\, h \in H\}$$



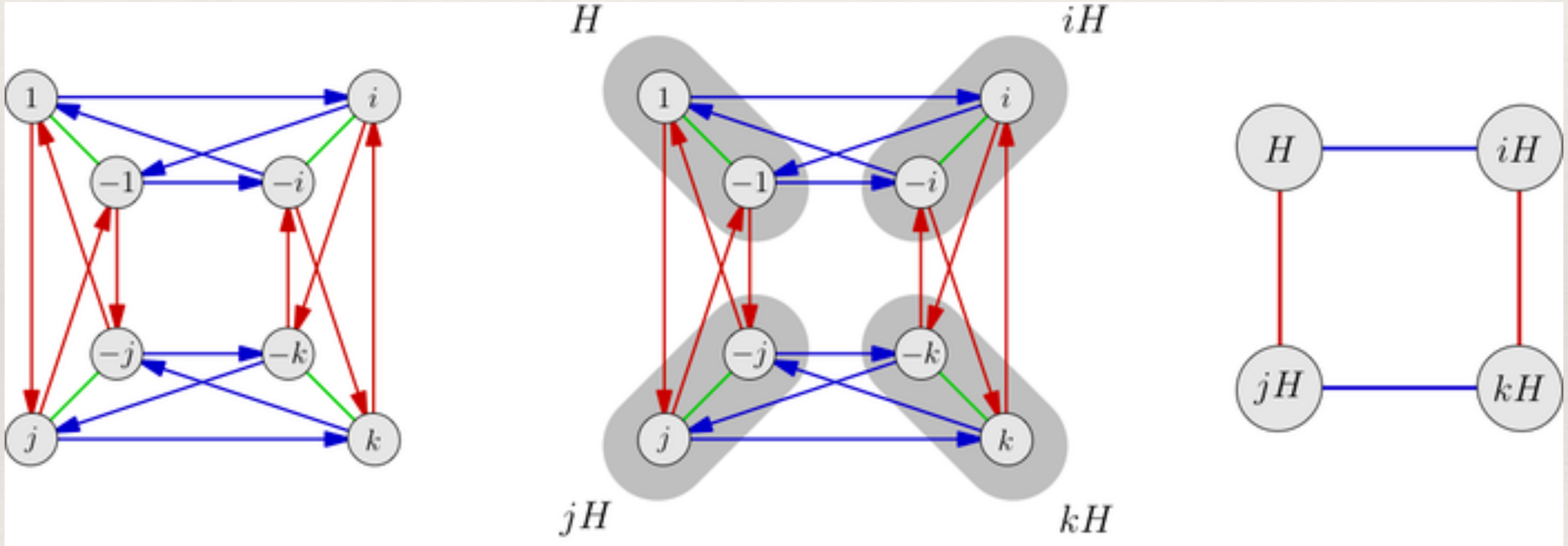
Question: when is a coset a subgroup?

Question: do the cosets always partition the group?

Quotients

The quotient of G by subgroup H is the set of cosets of H in G

$$G/H = \{gH \mid g \in G\}$$

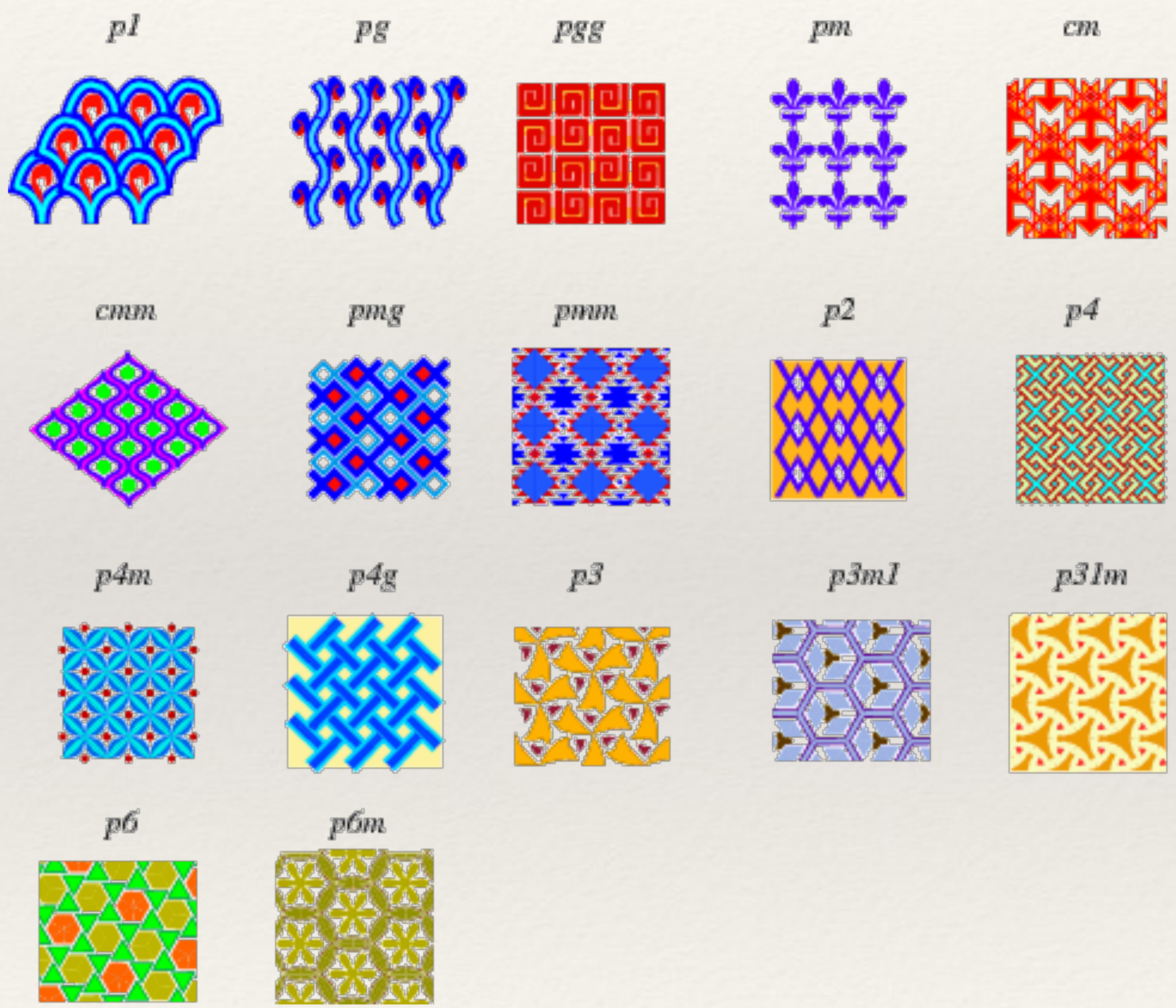


Q_4 organized by the subgroup $H = \langle -1 \rangle$

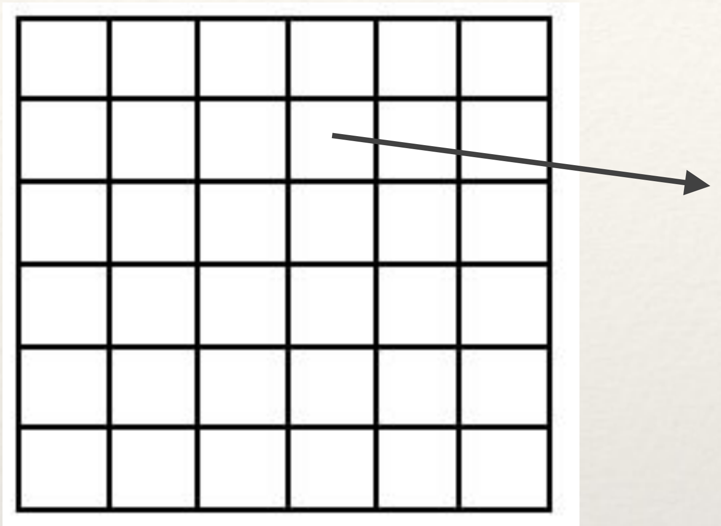
left cosets of H are near each other

collapse cosets into single nodes

Wallpaper Groups



The Groups p4 & p4m



$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

$$g(m,r,u,v) = \begin{bmatrix} (-1)^m & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(r\pi/2) & -\sin(r\pi/2) & 0 \\ \sin(r\pi/2) & \cos(r\pi/2) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & u \\ 0 & 1 & v \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} (-1)^m \cos(r\pi/2) & -(-1)^m \sin(r\pi/2) & u \\ \sin(r\pi/2) & \cos(r\pi/2) & v \\ 0 & 0 & 1 \end{bmatrix}$$

The Goups p6 & p6m

$$g(m,r,u,v) = \begin{bmatrix} (-1)^m & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(r\pi/3) & -\sin(r\pi/3) & 0 \\ \sin(r\pi/3) & \cos(r\pi/3) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & u \\ 0 & 1 & v \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} (-1)^m \cos(r\pi/3) & -(-1)^m \sin(r\pi/3) & u \\ \sin(r\pi/3) & \cos(r\pi/3) & v \\ 0 & 0 & 1 \end{bmatrix}$$

Group Equivariant CNNs

How to think about CNNs

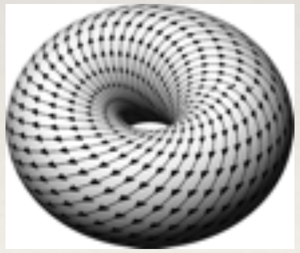
"A stack of feature maps is a 3D array"

"A stack of feature maps is a vector-valued function"

$$f^l : \mathbb{Z}^2 \rightarrow \mathbb{R}^{K_l}$$

"Mmm… Donuts"

"Genus one topological space"



G-Equivariant Correlation on Z^2

Standard correlation:
"translate canonical filter and compute inner product"

G-Correlation:

"*transform* canonical filter and compute inner product"

Translational Correlation

- Translation

$$[T_s f](x) = f(x - s)$$

- Correlation

$$[f \star \psi](s) = \sum_{x \in \mathbb{Z}^2} \sum_{k=1}^{K} f_k(x)[T_s \psi]_k(x)$$

- Equivariance

$$[T_s f] \star \psi = T_s[f \star \psi]$$

Group Correlation on Z^2

- Transformation

$$[T_g f](x) = f(g^{-1}x)$$

- G-Correlation

$$[f \star \psi](g) = \sum_{x \in \mathbb{Z}^2} \sum_{k=1}^{K} f_k(x)[T_g \psi]_k(x)$$

- Equivariance

$$[T_g f] \star \psi = T_g [f \star \psi]$$

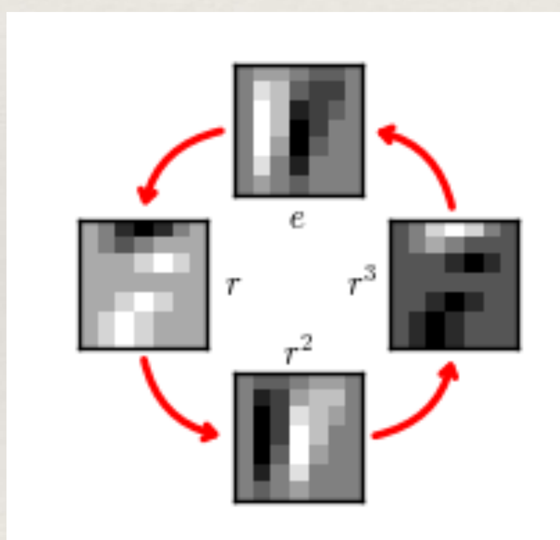
Feature Transformation Law (p4)

$$[T_g f] \star \psi = T_g [f \star \psi]$$

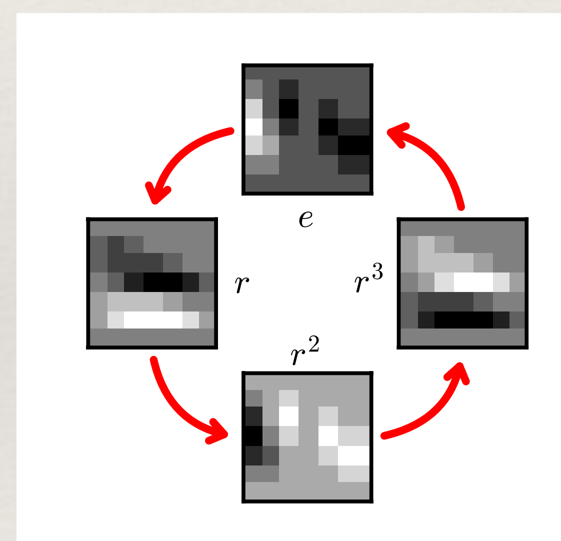
Transformation of 2D image

???

$T_r \Bigg($



$\Bigg) =$

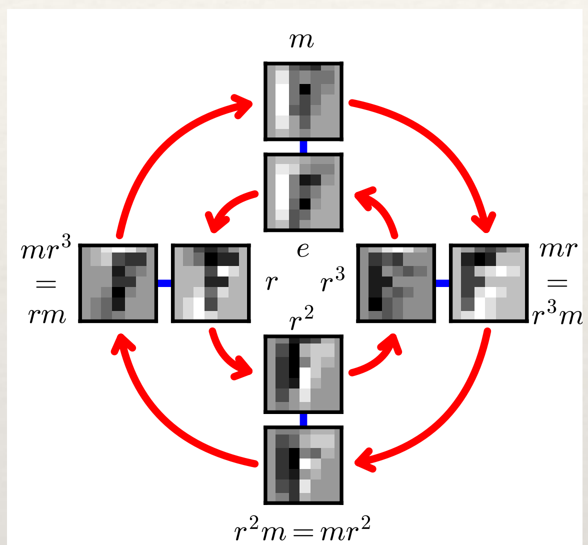


$$\begin{bmatrix} R(\theta') & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} R(\theta) & t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R(\theta - \theta') & R(-\theta')t \\ 0 & 1 \end{bmatrix}^{-1}$$

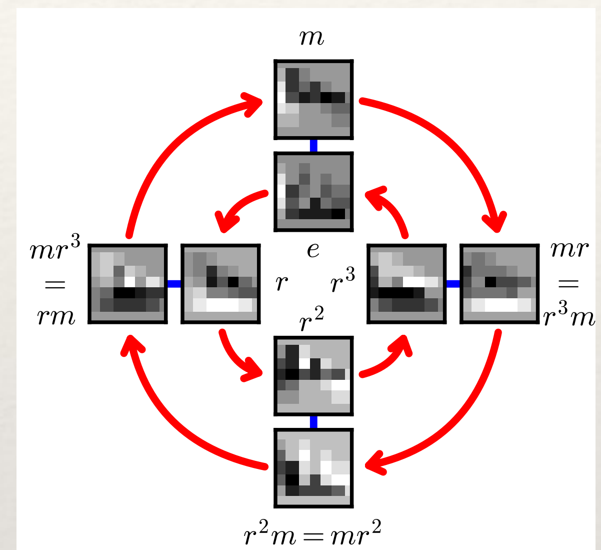
Feature Transformation Law (p4m)

$$[T_g f] \star \psi = T_g [f \star \psi]$$

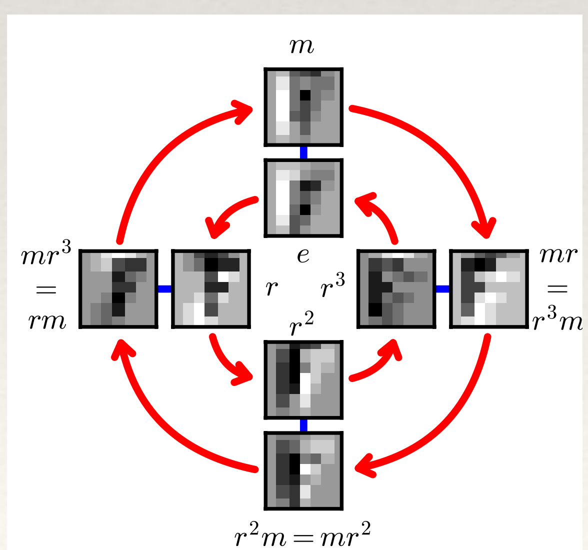
$T_r \Big($



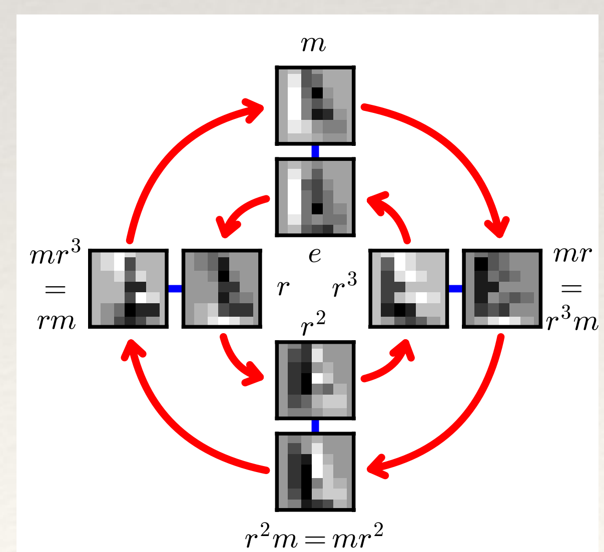
$\Big) =$



$T_m \Big($



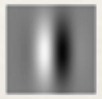
$\Big) =$



G-Conv is Non-Commutative

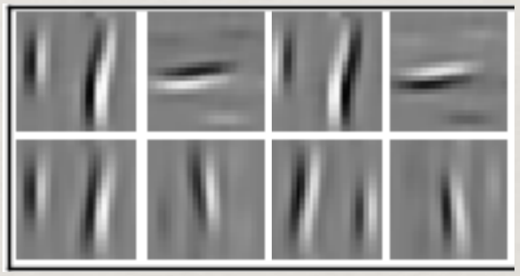
G = p4

f = 

psi = 

$f \star \psi$

$\psi \star f$



Same information content: $f \star \psi(g^{-1}) = \psi \star f(g)$

Group Correlation on G

- Transformation

$$[T_g f](h) = f(g^{-1}h)$$

- G-Correlation

$$[f \star \psi](g) = \sum_{h \in G} \sum_{k=1}^{K} f_k(h)[T_g \psi]_k(h)$$

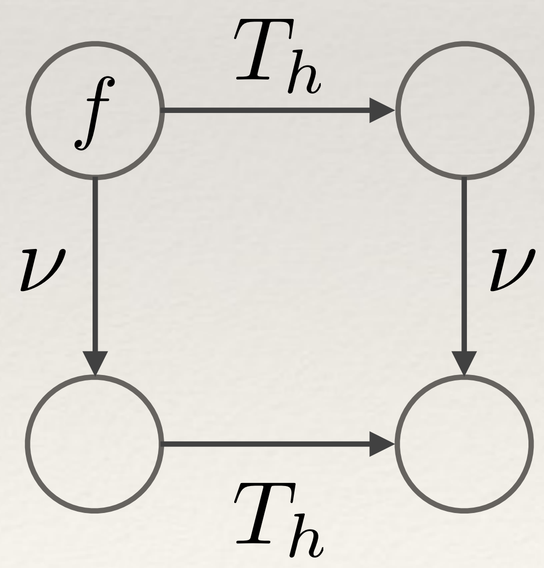
- Equivariance

$$[T_g f] \star \psi = T_g [f \star \psi]$$

Non-linearities: Equivariant

Function Composition Commutes with Domain Transformations

$$\nu \circ [T_h f] = \nu \circ (f \circ h^{-1}) = (\nu \circ f) \circ h^{-1} = T_h[\nu \circ f]$$



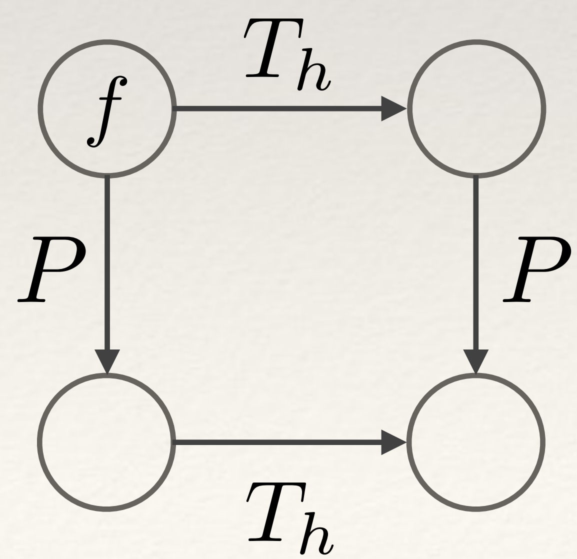
Strideless G-Pooling: Equivariant

Max-pool over neighborhood gU of g:

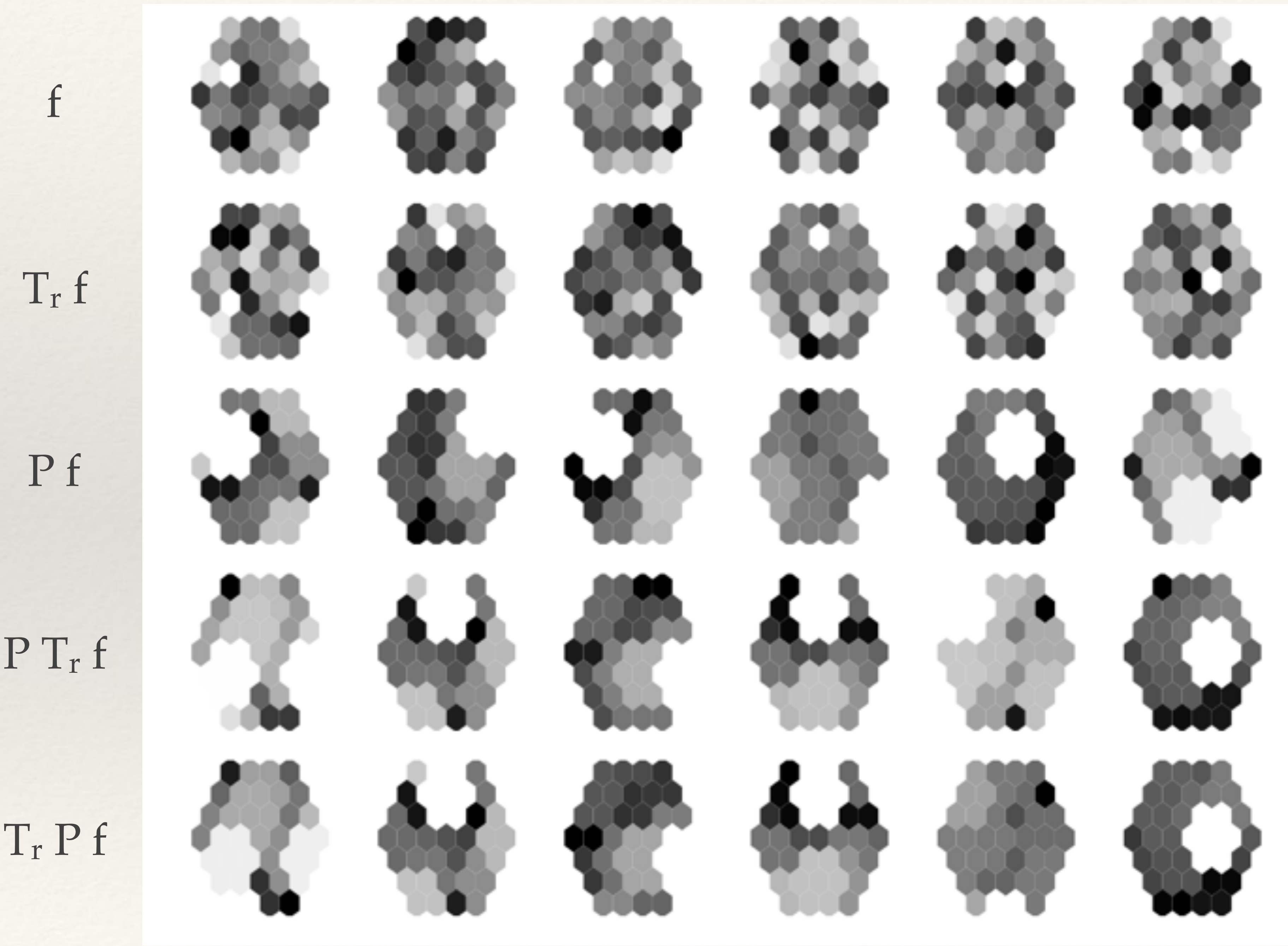
$$Pf(g) = \max_{k \in g \cdot U} f(k)$$

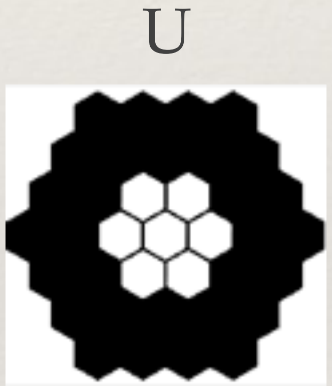
Pooling Operator Commutes with G-Action

$$PT_h = T_h P$$



Z^2 pooling of a p6 feature map

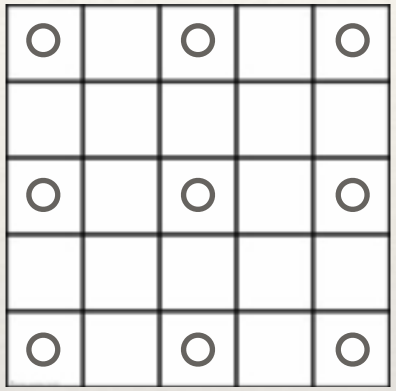


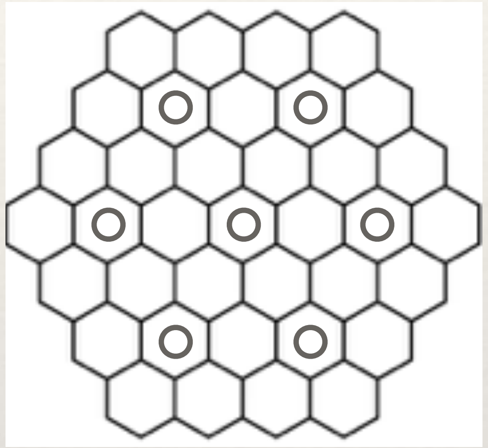


Subsampling

Subsample on a subgroup H of G

$$2\mathbb{Z}^2 \subset \mathbb{Z}^2$$





$$C^4 \ltimes 2\mathbb{Z}^2 \subset C^4 \ltimes \mathbb{Z}^2$$

$$C^4 \subset C^4 \ltimes \mathbb{Z}^2$$

$$\mathbb{Z}^2 \subset C^4 \ltimes \mathbb{Z}^2$$

Coset Pooling

Choose pooling domain U to be a subgroup H

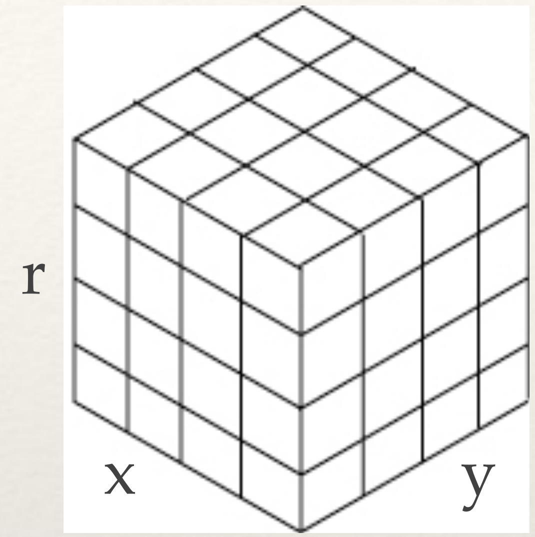
$$gH = \{gh \,|\, h \in H\}$$

Pooled feature map is constant on cosets

$$Pf(gh) = \max_{k \in ghH} f(k) = \max_{k \in gH} f(k)$$

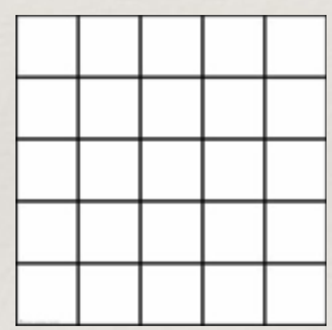
That is, the pooled feature map is a function on the quotient G / H

Example: p4 Coset Pooling



p4 feature map array

Pool over $U = C^4$

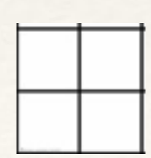


$p4/C^4 = Z^2$ feature map array

Pool over $U = \mathbb{Z}^2$

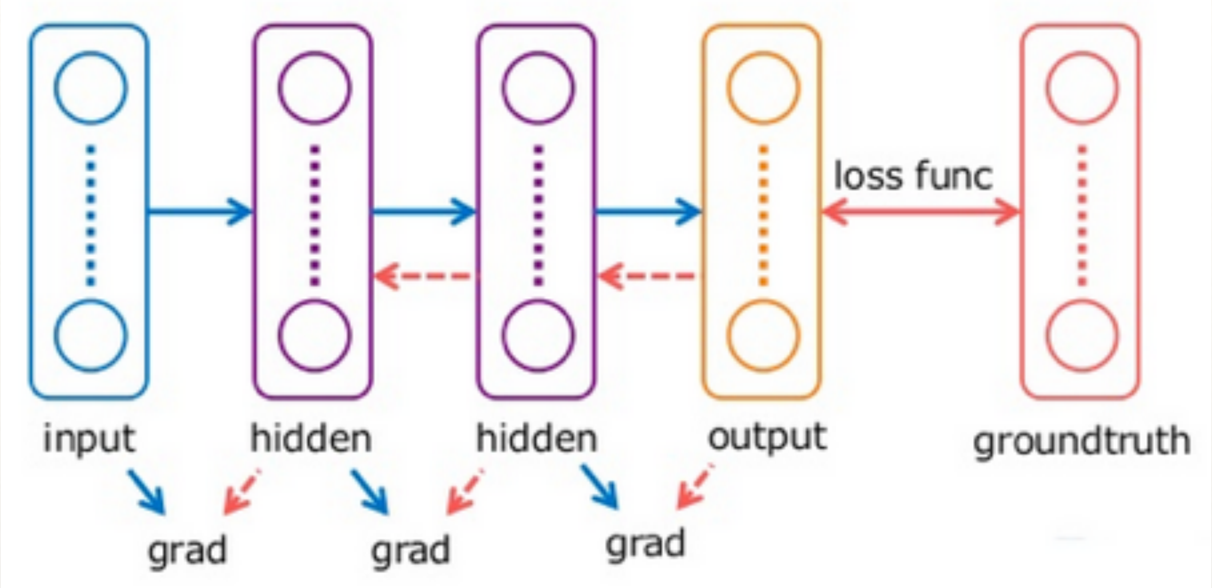
$p4/Z^2 = C^4$ feature map array

Pool over $U = C^4 \ltimes 2\mathbb{Z}^2$



$Z^2/2Z^2$ feature map array

Backprop



Input grad: involuted correlation

$$\frac{\delta L}{\delta f_j^{l-1}(k)} = \frac{\delta L}{\delta f^l} \star \psi_j^{l*}(k)$$

Filter grad: involuted convolution

$$\frac{\delta L}{\delta \psi_j^{li}(k)} = \frac{\delta L}{\delta f_i^l} * f_j^{l-1*}(k)$$

Algorithms:
Spatial & Spectral G-Convs

Naive Spatial Implementation

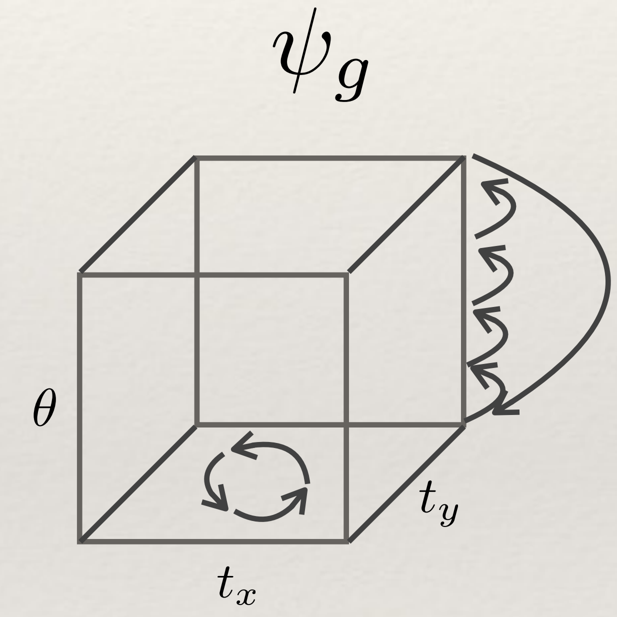
Compute: $r(g) = f * \psi(g)$

For each output channel j:

For each g in G-grid:

Warp filter: $\psi_g^j = T_g \psi^j$

Compute inner product:



$$r^j[g] = \langle f, \psi_g \rangle = \sum_i \sum_{t_x} \sum_{t_y} \sum_{\theta} f^i[t_x, t_y, \theta] \psi_g^{ji}[t_x, t_y, \theta]$$

Efficient Spatial Implementation

G-Correlation

$$[f \star \psi](g) = \sum_{h \in G} \sum_{k=1}^{K} f_k(h)[T_g \psi]_k(h)$$

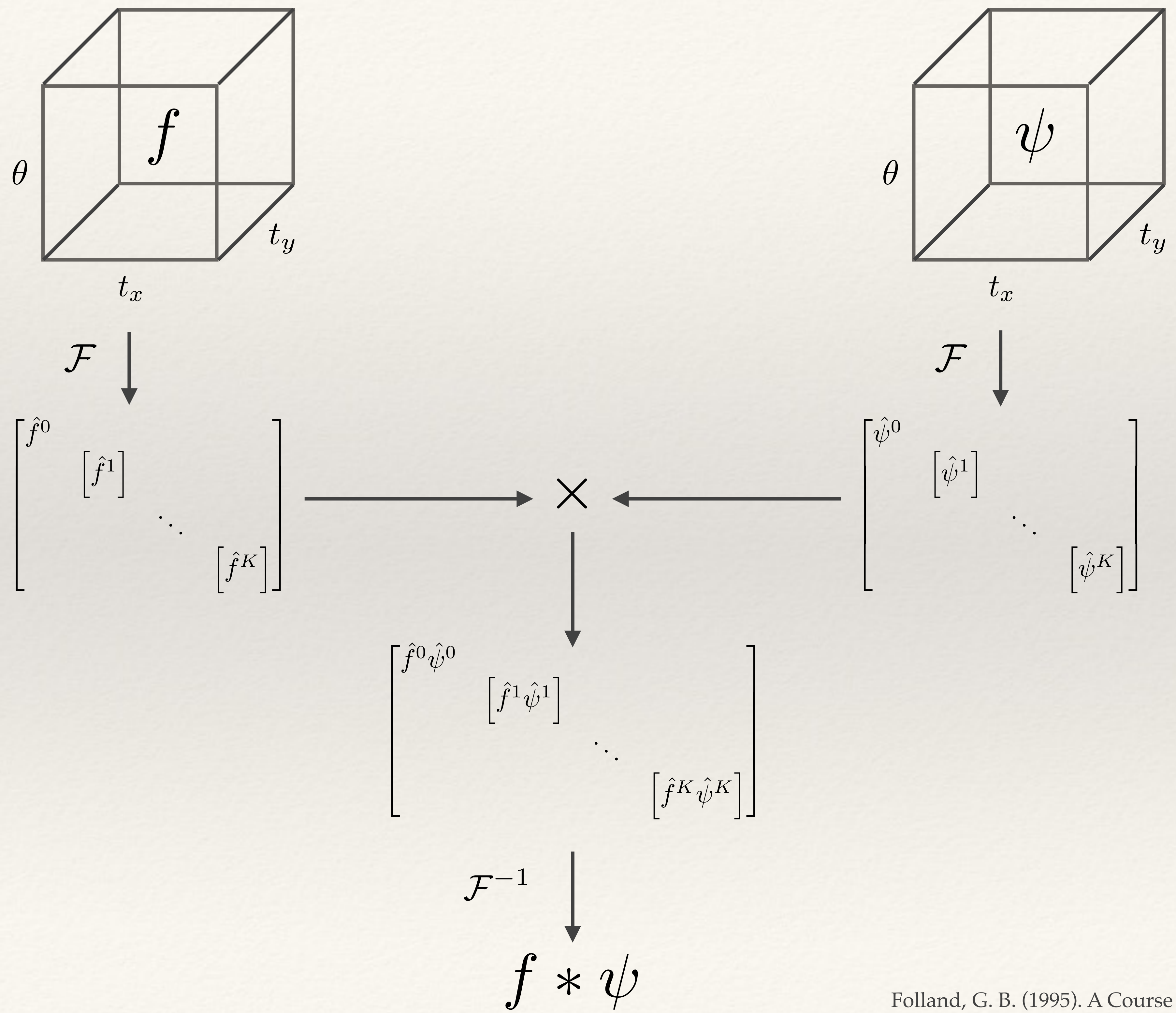
Decompose into translation and rotation:

$$T_g = T_{tr} = T_t T_r$$

To get output feature plane theta, do a planar convolution with rotated filter:

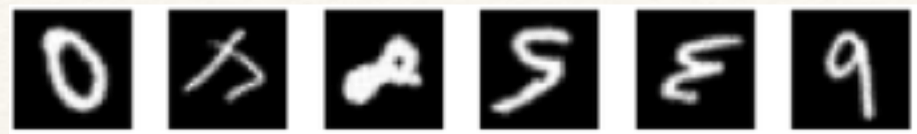
$$[f \star \psi](t, r) = \sum_{h} \sum_{k} f(h) L_t \psi_r(h)$$

Spectral G-Convolution



Folland, G. B. (1995). A Course in Abstract Harmonic Analysis.

Results: Rotated MNIST



Z2CNN	P4CNNRotPool	P4CNN
C(1, 16, 7)	P4C1(1, 16, 7)	P4C1(1, 10, 3)
MP(3, 2)	MPR + MP(3, 2)	P4C(10, 10, 3)
C(16, 16, 5)	P4C1(16, 16, 5)	MP(2, 2)
MP(3, 2)	MPR + MP(3, 2)	P4C(10, 10, 3)
C(16, 32, 5)	P4C1(16, 32, 5)	P4C(10, 10, 3)
MP(3, 2)	MPR + MP(3, 2)	P4C(10, 10, 3)
C(32, 10, 3)	C(32, 10, 3)	P4C(10, 10, 3)

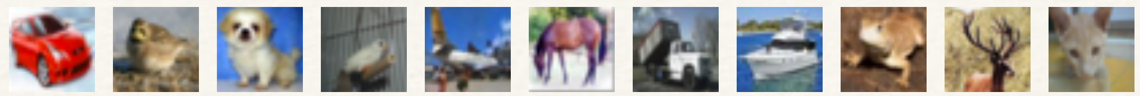
Table 1. Rotated MNIST architectures.
Legend: C(in, out, sz) denotes a convolution layer with given number of input channels, output channels and kernel size, P4C1(in, out, sz) denotes first-layer P4 convolution while P4C denotes full P4 convolutions, MP(ksize, stride) denotes max pooling and MPR denotes max pooling over all 4 rotation angles.

SVM[1]	NNet[1]	DBN[1]	RC-RBM[2]	Z2CNN	P4CNNRP	P4CNN
10,38	17,62	12,11	3,98	5,68	3,90	**2.51**

[1] Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007) An empirical evaluation of deep architectures on problems with many factors of variation. (ICML)

[2] Schmidt, U., & Roth, S. (2012). Learning rotation-aware features: From invariant priors to equivariant descriptors. (CVPR)

CIFAR-10



All-CNN

3 x 3 conv. 96 ReLU
3 x 3 conv. 96 ReLU
3 x 3 conv. 96 ReLU (stride 2)
3 x 3 conv. 192 ReLU
3 x 3 conv. 192 ReLU
3 x 3 conv. 192 ReLU (stride 2)
3 x 3 conv. 192 ReLU
1 x 1 conv. 192 ReLU
1 x 1 conv. 10 ReLU
Global averaging
softmax

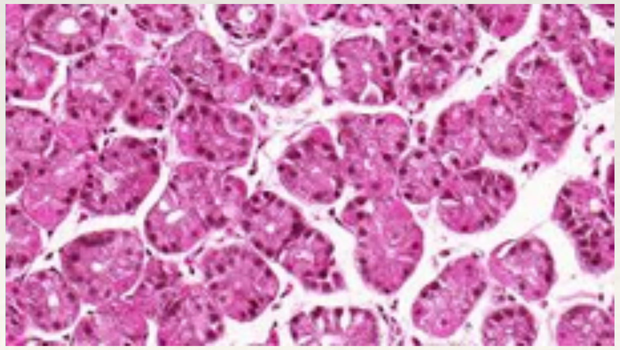
P4-All-CNN

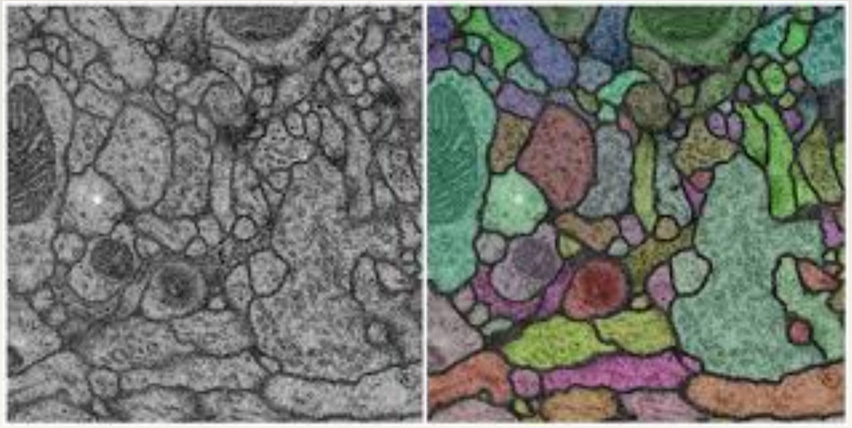
Replace conv by p4-conv,
halve number of filters

Results

Model	Test Error (%)
Maxout (Goodfellow et al., 2013)	11.68
NiN (Lin et al., 2013)	10.41
DSN (Lee et al., 2015)	9.69
All-CNN (our baseline)	9.82
All-CNN (Springenberg et al., 2015)	9.07
All-CNN-BN (added batchnorm)	9.44
P4-All-CNN (ours)	8.84
Elu (Clevert et al., 2015)	6.55

Work in Progress..





Projects / Theses / Internships

taco.cohen@gmail.com

Questions?