# MATHEMATICS OF NEURAL NETWORKS

## OUTLINE

- GROUPS
- ACTIONS/LINEAR REP.
- EQUIVARIANCE/INVARIANCE
- DATA AUGMENTATION/DAC
- ENSEMBLING/REYNOLD AVERAGING
- IRREDUCIBLE REPRESENTATIONS/ SCHUR'S LEMMA
- EQUIVARIANT MAPPINGS.
- BEYOND IMAGES
    - GRAPHS
    - POINT CLOUDS
    - ETC.

## Groups

A GROUP G VERIFIES THE FOLLOWING AXIOMS:

$\exists e \in G : \forall g \in G \quad g \circ e = g$ identity element.

$\forall g, s \in G \quad g \circ s \in G$

$\forall g \in G \ \exists s \in G \quad g \circ s = e$

EXAMPLES:

- $\mathbb{Z}_n = \{e, g, g^2, \dots, g^{n-1}\}$

  $g^k \circ g^{n-k} = e$

- $(\mathbb{R}_+, \times)$

GROUP ACTION : FOR A VECTOR SPACE $V$ (e.g. $\mathbb{R}^n$)

$$T : G \times V \to V$$ IS A GROUP ACTION

iff it verifies

$$T_g \circ T_s = T_{g \circ s}$$

$T_e v = v$ identity mapping $\forall v \in V$

LINEAR REPRESENTATIONS = LINEAR GROUP ACTIONS

$$T_g(\alpha v + \beta u) = \alpha T_g v + \beta T_g u$$

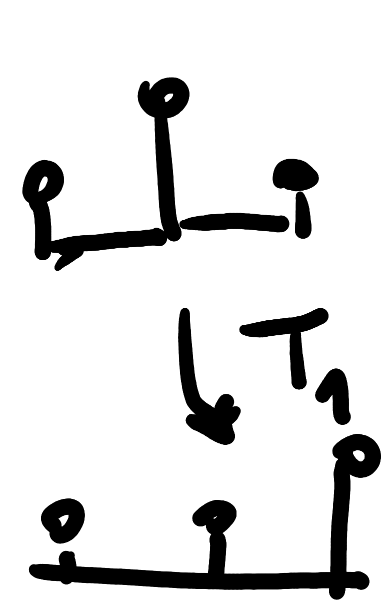
$$\alpha, \beta \in \mathbb{R} \quad u, v \in V$$

EXAMPLE $V = \mathbb{R}^n \Rightarrow T_g \in \mathbb{R}^{n \times n} \quad \forall g \in G$

INVERTIBLE MATRICES

$$T_e = I_n \quad T_g^{-1} = T_{g^{-1}}$$

- SHIFTS OF A DISCRETE TIME SERIES ($n=3$)




$$\{T_1, T_2, I\} \qquad T_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbb{Z}_3$$

$$T_2 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

- TRIVIAL REPRESENTATION

$$T_g = I \quad \forall g \in G$$ IS A VALID GROUP ACTION

- REGULAR REPRESENTATION

$$V = \mathbb{R}^{|G|}$$

$$\underbrace{T_g}_{\text{PERMUTATION MATRICES}} e_s = e_{g \circ s}$$

$$e_s = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \rightarrow s_{\text{TH}} \text{ POSITION}$$

---

<u>EQUIVARIANCE:</u> A FUNCTION $f : \mathbb{R}^m \to \mathbb{R}^p$ IS EQUIVARIANT iif

$$f(T_g x) = \tilde{T}_g f(x) \quad \forall g \in G, \ \forall x \in \mathbb{R}^m$$

if $\underbrace{T \text{ AND } \tilde{T}}_{\text{MIGHT BE DIFFERENT!}}$ ARE TWO VALID GROUP ACTIONS (LINEAR OR NON LINEAR)

<u>EXAMPLE</u>

$f(x) = x$ is $G$ EQUIVARIANT for ANY GROUP.

# Properties of linear reps.

$T_g$ : action on $\mathbb{R}^n$

$\tilde{T}_g$ " " $\mathbb{R}^m$

1) $\begin{bmatrix} T_g & \\ & \tilde{T}_g \end{bmatrix}$ linear rep on $\mathbb{R}^{n+m}$

2) $T_g \otimes \tilde{T}_g$ linear rep on $\mathbb{R}^{nm}$

kronecker product

3) for invertible $A \in \mathbb{R}^{n \times n}$

$A T_g A^{-1}$ linear rep on $\mathbb{R}^n$

4) $T_g$ linear rep of $G$

$\tilde{T}_{g'}$ linear rep of $G'$

$\Rightarrow T_g \tilde{T}_{g'}$ linear rep of $G \times G'$

# INVARIANCE: $f$ is INVARIANT iff

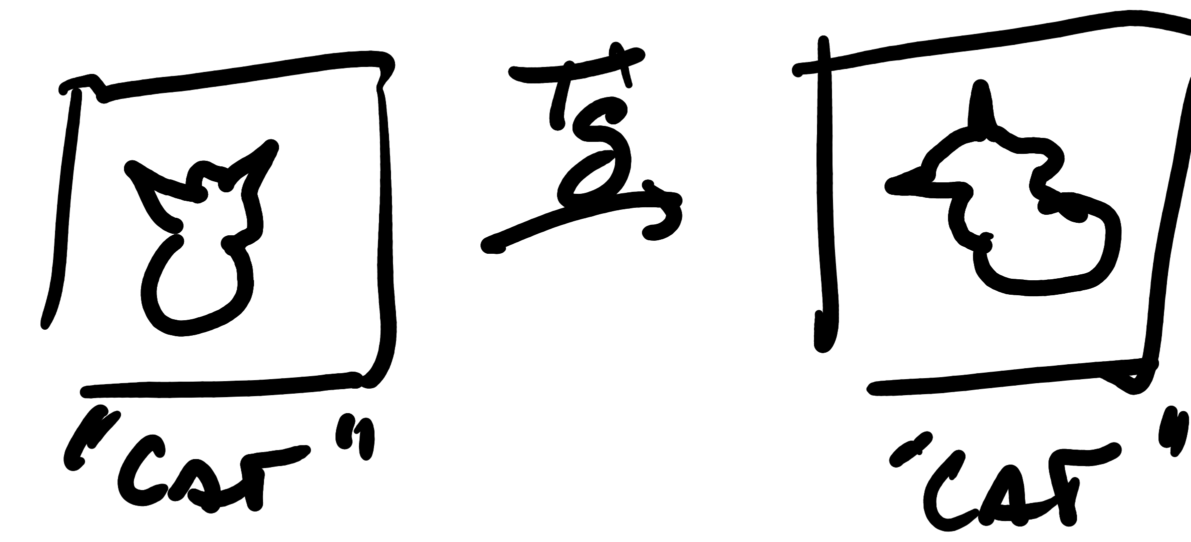
$$f(T_g x) = x \quad \forall g \in G, \ \forall x \in \mathbb{R}^n$$

FOR A VALID ACTION $T$

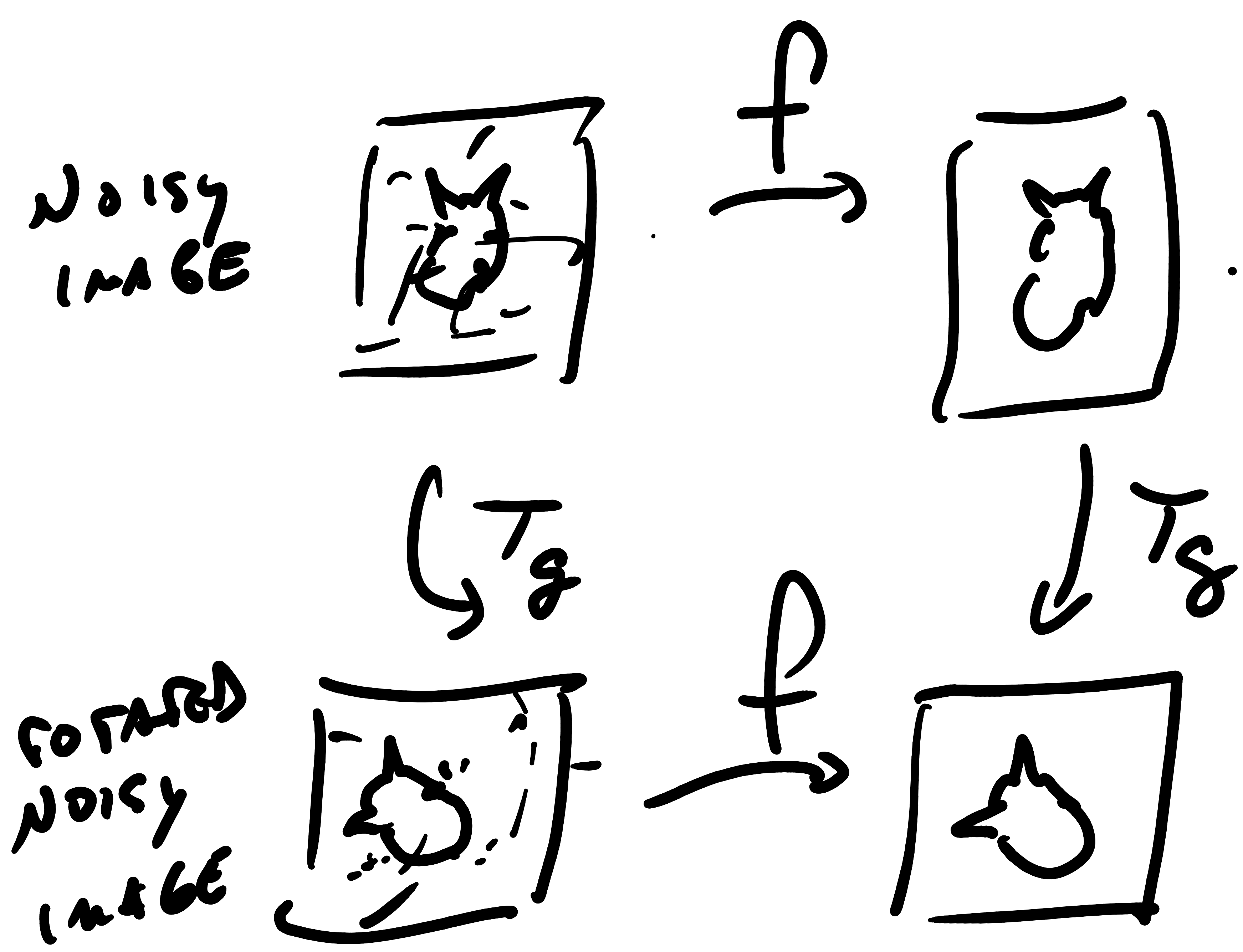
**NOTE**: THIS IS A SPECIAL CASE OF EQUIVARIANCE WHERE $\tilde{T}$ IS THE TRIVIAL REPRESENTATION

## EXAMPLES IN ML

IN CLASSIFICATION WE LOOK FOR INVARIANT FUNCTIONS




IN REGRESSION WE LOOK FOR EQUIVARIANT FUNCTIONS



# DATA AUGMENTATION

$$\arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i)$$

CLASSICAL EMPIRICAL RISK. MINIMIZATION

$$\arg\min_\theta \sum_{i=1}^{N} \underbrace{\sum_{g\in G}}_{} \mathcal{L}(f_\theta(T_g x_i), \tilde{T}_g y_i)$$

GENERALLY HANDLED IN A STOCHASTIC WAY

- DATA AUG. CONSISTENCY

$$\arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i) + \underbrace{\sum_{g\in G} \mathcal{L}(f_\theta(x_i), f_\theta(T_g x_i))}_{}$$

OFTEN USED FOR SELF-SUPERVISED LEARNING?

# GROUP AVERAGING / ENSEMBLING

G-EQUIVARIANT $f: \mathbb{R}^n \to \mathbb{R}^m$

$$f^E(x) = \frac{1}{|G|} \sum_{g \in G} T_g^{-1} f(T_g x)$$

INVARIANT $f: \mathbb{R}^n \to \mathbb{R}$

$$f^I(x) = \frac{1}{|G|} \sum_{g \in G} f(T_g x)$$

- TOO EXPENSIVE FOR LARGE GROUPS...
- STILL IT IS POPULAR IN COMPUTER VISION FOR 90° ROTATIONS + VERTICAL AND HORIZ. FLIPS.

- ALSO CALLED 'REYNOLDS AVERAGING'.

# EQUIVARIANT NETWORKS

→ CONSTRAIN WEIGHTS $\theta$ SUCH THAT

$f_\theta$ IS $G$ EQUIVARIANT FOR ALL $\theta$.

→ DEEP NETWORKS :

- MAINTAIN EQUIVARIANCE UNTIL THE LAST LAYER + INVARIANCE IN THE FINAL LAYERS.

## CNN

CONVOLUTION LAYERS + POINTWISE NON-LINEARITY → TRANSLATION EQUIVARIANT

POOLING LAYERS → TRANSLATION INVARIANT

$$x^{\ell+1} = \phi(W x^{\ell})$$

## EQUIVARIANT LINEAR LAYER.

WHAT IS THE SET OF EQUIVARIANT MATRICES?

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_c \end{bmatrix}$$

$$S = \{ W \in \mathbb{R}^{m_o \times m_i} : \quad \tilde{T}_g W = W T_g \quad \forall g \in G \}$$

EXAMPLE $\tilde{T}_g = T_g$ = Group of SHIFTS.
$m_o = m_i = m$ ON OF $m$ ELEMENTS

$$T_g W = T_g W \iff W = \mathrm{circ}(\underset{\in \mathbb{R}^m}{w})$$

$$\dim(S) = m$$

# IRREDUCIBLE REPRESENTATIONS:

- A LINEAR REP. IS IRREDUCIBLE IIF IT CANNOT BE DECOMPOSED AS

$$T_g = A \begin{bmatrix} \rho_g^1 & \\ & \rho_g^2 \end{bmatrix} A^{-1} \quad \forall g \in G$$

WHERE $\rho^1$ AND $\rho^2$ ARE VALID LINEAR REP.

<u>e.g.</u>

ACTION OF $\mathbb{Z}_2$

$$T_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad T_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$T_g = A \begin{bmatrix} 1 & \\ & (-1)^g \end{bmatrix} A^{-1} \quad g = \{0, 1\}$$

CAN BE REDUCED TO

$$T_g = A \begin{bmatrix} \rho^1(g) & \\ & \rho^2(g) \end{bmatrix} A^{-1}$$

WHERE $\begin{cases} \rho_1 = 1 \quad \forall g \text{ TRIVIAL REP} \\ \rho^2(g) = (-1)^g \end{cases}$

**THEOREM**: A (FINITE) COMPACT GROUP $G$ HAS $K$ **DISTINCT** IRREDUCIBLE REPRESENTATIONS

$$\{\rho^1, \dots, \rho^K\} \text{ SUCH THAT}$$

$$\sum_{i=1}^{K} (\dim \rho^i)^2 = |G|$$

**e.g.** $\mathbb{Z}_2$ HAS

$$\rho^1(g) = 1$$
$$\rho^2(g) = (-1)^g$$

$$\dim \rho^1 = 1 \qquad |G| = 2 = 1^2 + 1^2$$
$$\dim \rho^2 = 1$$

**THEOREM**: ANY LINEAR REP. CAN BE DECOMPOSED IN BLOCK DIAGONAL FORM:

$$T_g = A \begin{bmatrix} \left.\begin{matrix} \rho_1 & & \\ & \ddots & \\ & & \rho_1 \end{matrix}\right\} m_1 \text{ TIMES} & & \\ & m_2 \text{ TIMES} \left\{\rho_2\right. & \\ & & \ddots \\ & & & \rho_K \end{bmatrix} A$$

WHERE $A$ IS AN ORTHOGONAL BASIS

WHERE $\rho_j$ IS REPEATED $m_j$ TIMES SUCH THAT $\sum m_j \dim \rho_j = n$

EXAMPLE : SHIFT MATRICES (ACTION OF $\mathbb{Z}_m$)

$$T_g = F \begin{bmatrix} \rho_g^1 & & \\ & \ddots & \\ & & \rho_g^m \end{bmatrix} F^*$$

→ $F$ IS THE DISCRETE FOURIER BASIS.

→ $\rho_g^j = e^{-\frac{i2\pi j}{m} g}$ ARE THE (1-DIM.) IRREPS OF $\mathbb{Z}_m$

THEOREM: EQUIVARIANT MATRICES

$$S = \{ W : \tilde{T}_g W = W T_g \} \text{ WITH } T_g = A^{-1} \begin{bmatrix} \rho_1 & & \\ & \ddots & \\ & & \rho_K \end{bmatrix} A$$

$$\tilde{T}_g = B^{-1} \begin{bmatrix} \rho_1 & & \\ & \ddots & \\ & & \rho_K \end{bmatrix} B$$

if AND ONLY IF WE CAN WRITE $W$ AS

$$W = A^{-1} \begin{bmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_K \end{bmatrix} B$$

WHERE $\psi_j \in \mathbb{C}^{\tilde{m}_j \dim \rho_j \times m_j \dim \rho_j}$

LINEAR SUBSPACE OF $\mathbb{R}^{m_o \times m_i}$

$\dim S =$ DEGREES OF FREEDOM $= \sum_{j=1}^{K} \tilde{m}_j m_j \dim \rho_j^2 \leq m_i m_o$

<u>EXAMPLE</u>: $T_g$ AND $\tilde{T}_g$ ARE SHIFTS IN $\mathbb{R}^n$

$$W = F \begin{bmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_n \end{bmatrix} F^*$$

$= \mathrm{circ}(\hat{\omega})$ CIRCULANT MATRICES!

$n$ DEGREES OF FREEDOM $\ll n^2$!

$\Rightarrow$ RESTRICTING THE NETWORK TO HAVE G-EQUIVARIANT LAYERS REDUCES THE NUMBER OF LEARNABLE PARAMETERS.

$\Rightarrow$ AN EQUIVARIANT NN JUST REQUIRES DEFINING THE IRREPS AND MULTIPLICITIES AT EACH LAYER

$\Rightarrow$ $W^E = \frac{1}{|G|} \sum_g T_g W T_g^{-1}$ is JUST AN ORTHOGONAL PROJECTION ONTO $S$

# NON-LINEARITIES:

- IF THE GROUP ACTION AT LAYER L IS REPRESENTED BY PERMUTATION MATRICES:
  (FOR EXAMPLE, THE REGULAR REPRESENTATION)

$$[T_g x]_s = x_{g\cdot s}$$

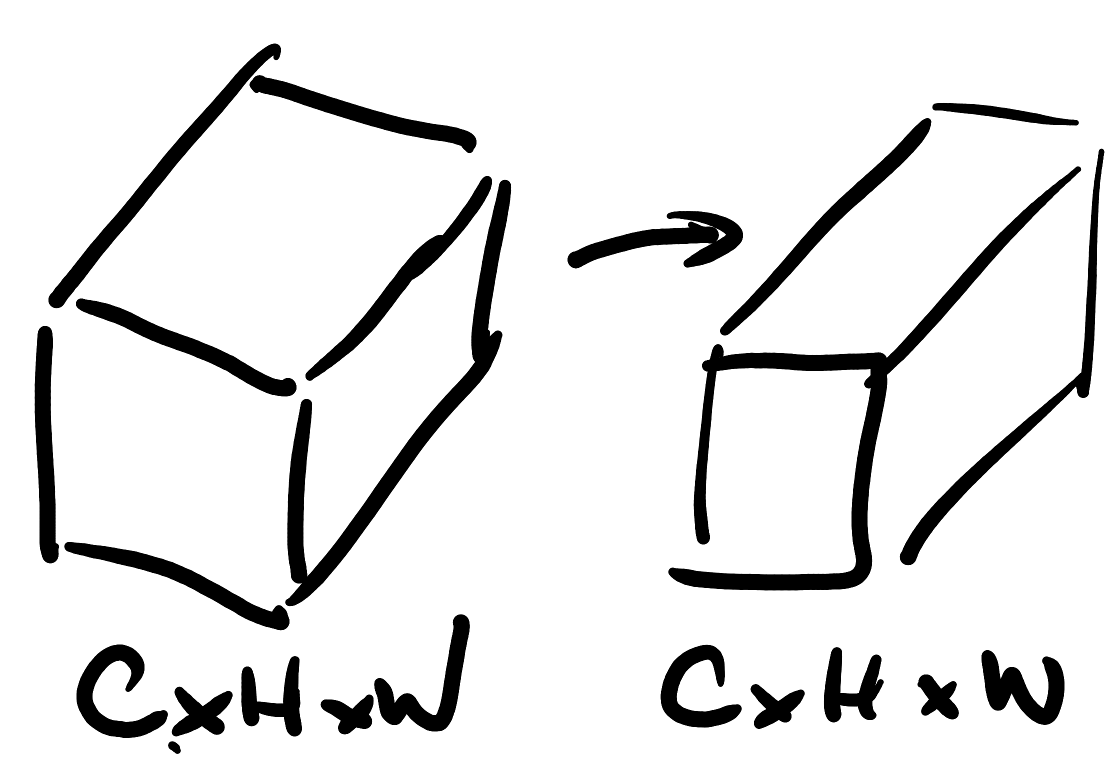
AN ELEMENTWISE NON-LINEARITY IS EQUIVARIANT:
$\phi : \mathbb{R} \to \mathbb{R}$

$$\phi(T_g x) = T_g \phi(x)$$

## EXAMPLE CNNs

$$x^\ell = \phi(W^\ell x^{\ell-1})$$

$$x^\ell = \begin{bmatrix} x_1^\ell \\ \vdots \\ x_C^\ell \end{bmatrix} \begin{array}{l} \to \text{CHANNEL } 1 \\ \\ \to \text{CHANNEL } C \end{array}$$




$$T_g^\ell = \begin{bmatrix} T_g & & \\ & \ddots & \\ & & T_g \end{bmatrix}$$

$T_g \to$ SHIFT MATRICES.

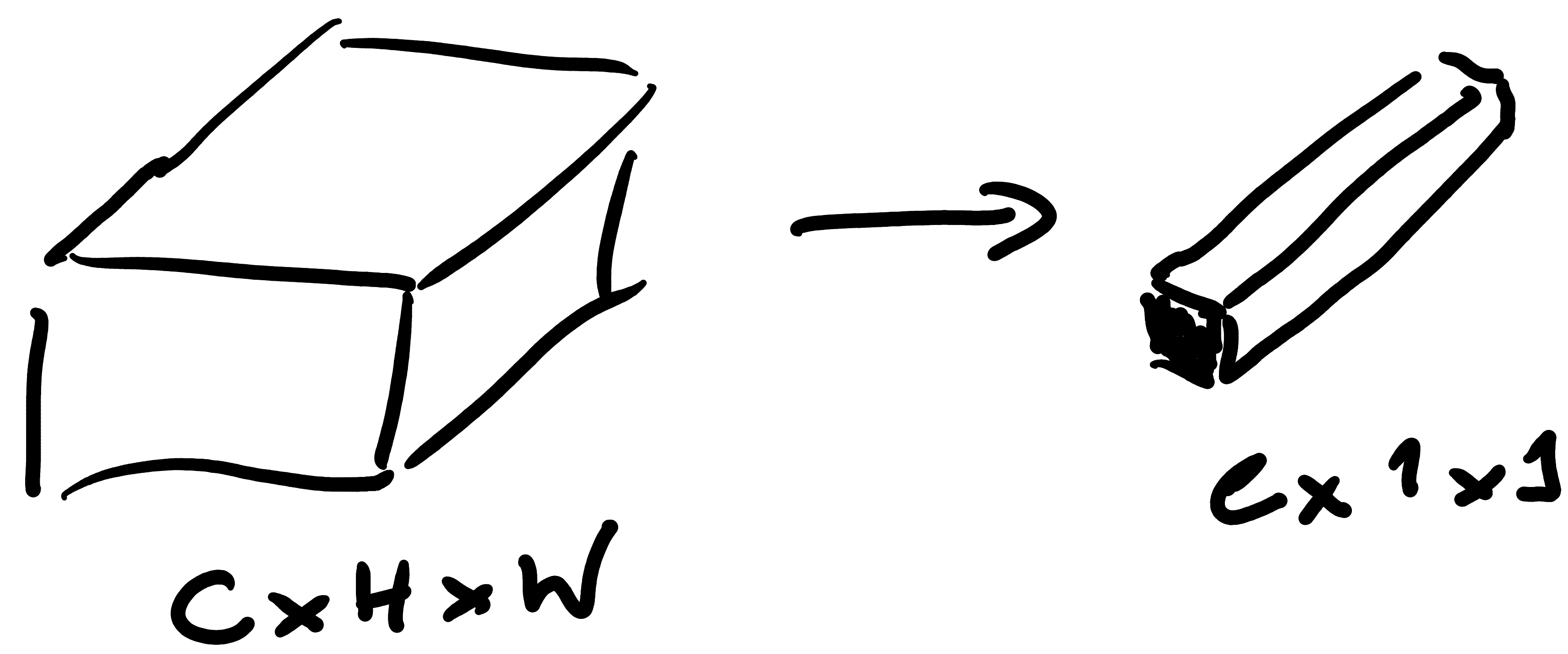
$$W^\ell = \begin{bmatrix} \mathrm{circ}\, W_{11} & \cdots & \mathrm{circ}\, W_{1C} \\ \vdots & \ddots & \vdots \\ \mathrm{circ}\, W_{C1} & \cdots & \mathrm{circ}\, W_{CC} \end{bmatrix}$$

INVARIANT LAYERS

$\hookrightarrow$ POOLING

OUTPUT ACTION = TRIVIAL ACTION

$$\underbrace{x^{\ell}}_{\in \mathbb{R}^{c}} = [1, \dots, 1] \, \underbrace{x^{\ell-1}}_{\in \mathbb{R}^{mc}}$$

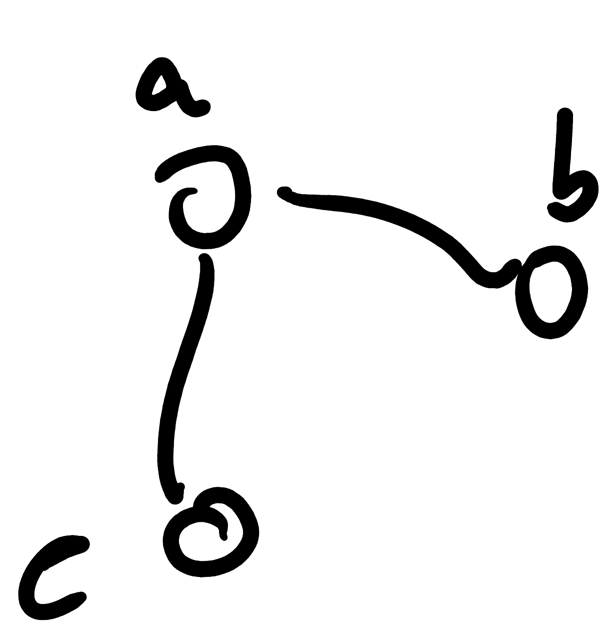



EXAMPLES:

- STEERABLE CNNs : EQUIVARIANCE TO ROTATIONS 90° AND HORIZONTAL AND VERTICAL FLIPS. (HOMEWORK).

- GRAPH NEURAL NETWORKS

GRAPHS ARE REPRESENTED BY ADJACENCY MATRICES




$$A = \begin{array}{c c} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

$$\tilde{A} = \begin{array}{c c} & \begin{matrix} b & c & a \end{matrix} \\ \begin{matrix} b \\ c \\ a \end{matrix} & \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \end{array}$$

THEY REPRESENT THE SAME GRAPH

ANY PERMUTATION MATRIX $T_g$ $g \in S_n$ DEFINES THE SAME GRAPH.

$$A' = T_g A T_g^T$$

GRAPH CLASSIFICATION PROBLEM:

LEARN $f(A) =$ GRAPH LABEL.

$f$ CAN BE DEFINED VIA PERMUTATION INVARIANT LAYERS.

WE CAN WRITE THE PROBLEM IN VECTOR FORM

$$\underbrace{\mathrm{vec}(A')}_{A \in \mathbb{R}^{n^2}} = \underbrace{T_g \otimes T_g}_{\tilde{T}_g \text{ GROUP ACTION ON } \mathbb{R}^{n^2}} \underbrace{\mathrm{vec}(A)}_{a \in \mathbb{R}^{n^2}}$$

$$\Leftrightarrow \quad a' = \tilde{T}_g a$$

SET OF PERMUTATION INVARIANT FUNCTIONS

## POINT CLOUDS

$$x = \begin{bmatrix} p_1 \\ \vdots \\ \vdots \\ p_m \end{bmatrix}$$

SET OF POINTS $p_i \in \mathbb{R}^3$

$x \in \mathbb{R}^{3m}$

GROUP ACTIONS:

- PERMUTATIONS OF POINTS $S_m$
- GLOBAL ROTATION + TRANSLATION OF POINTS.

$$SO(3) \quad (\mathbb{R}^3, +)$$

THIS IDEAS CAN BE EXTENDED TO ANY TYPE OF DATA AS LONG AS WE KNOW WHICH

EXAMPLES

- DATA ON A SPHERE (e.g. GLOBAL CLIMATE DATA)
- SETS
- TIME SERIES
- DATA ON GENERAL MANIFOLDS.