On training and application of equivariant neural networks.

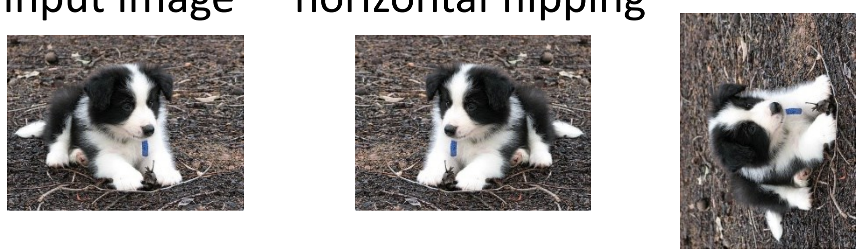
The University of Tokyo & RIKEN AIP

Mukuta Yusuke

Invariance in pattern recognition

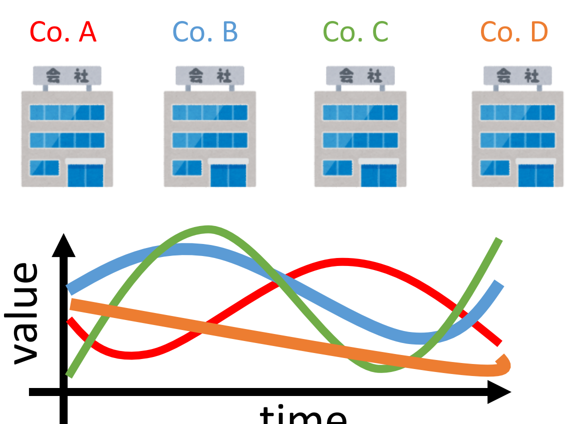
- The input signal for recognition often has invariance and equivariance with respect to transformations.

Image rotation and flip



'dog'

Permutation of multi-variate time series



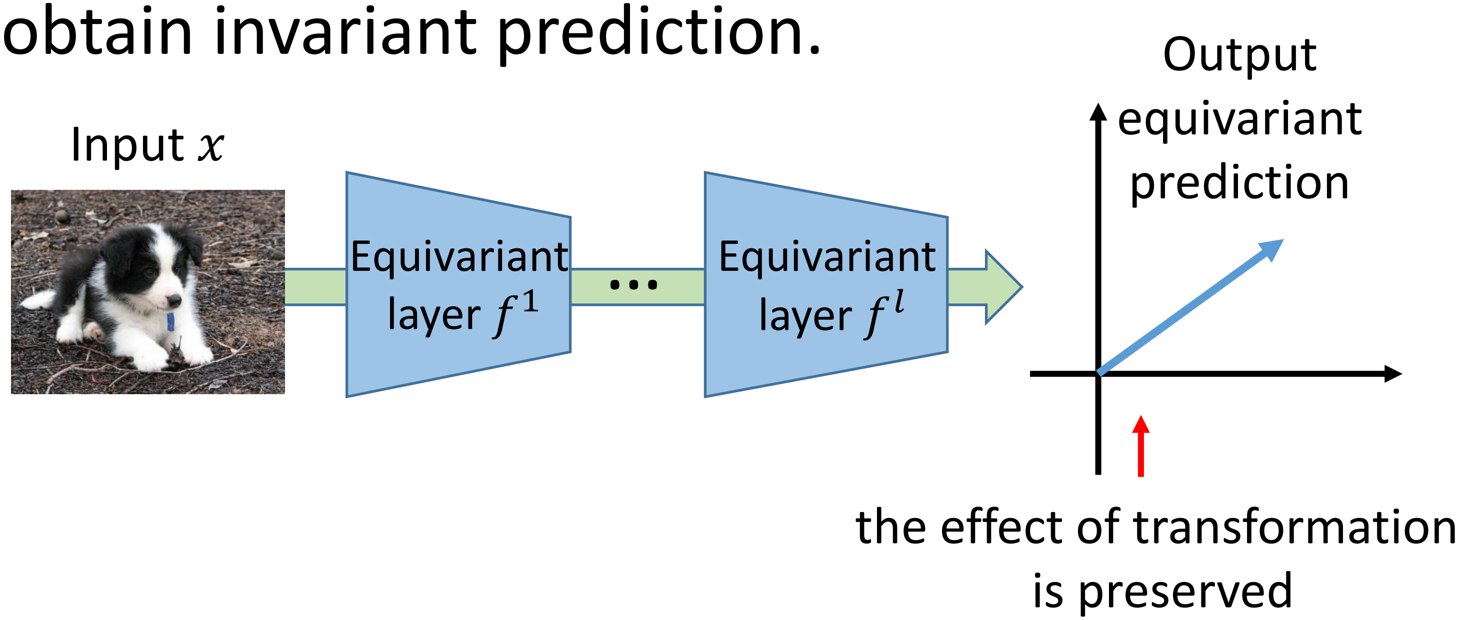
$$\begin{bmatrix} A_t \\ B_t \\ C_t \\ D_t \end{bmatrix} \xrightarrow{\text{Predictor } f} \begin{bmatrix} A_{t+1} \\ B_{t+1} \\ C_{t+1} \\ D_{t+1} \end{bmatrix}$$

equivalent

$$\begin{bmatrix} B_t \\ C_t \\ A_t \\ D_t \end{bmatrix} \xrightarrow{\text{Predictor } f} \begin{bmatrix} B_{t+1} \\ C_{t+1} \\ A_{t+1} \\ D_{t+1} \end{bmatrix}$$

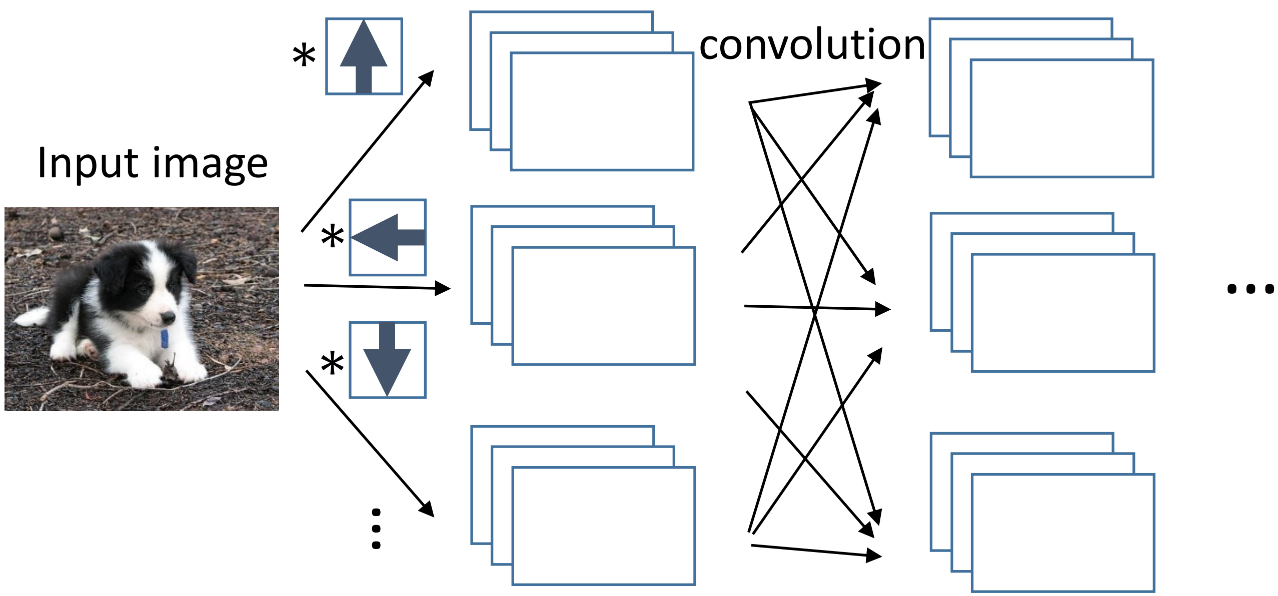
Equivariant Neural Network

- Equivariant Neural Network constructs the network by composing multiple equivariant layers.
- Given transformation T, equivariant layer f is a layer that commutes with T such that
  $$f\big(T_{in}(x)\big) = T_{out}(f(x)).$$
- We apply pooling with respect to transformations to obtain invariant prediction.



Group Equivariant Convolutional Networks [Cohen & Welling, 2016]

- Apply all the transformations to the convolutional filter and then apply convolution to the image.
- Then image transformations results in the permutation of the filter response (equivariant).

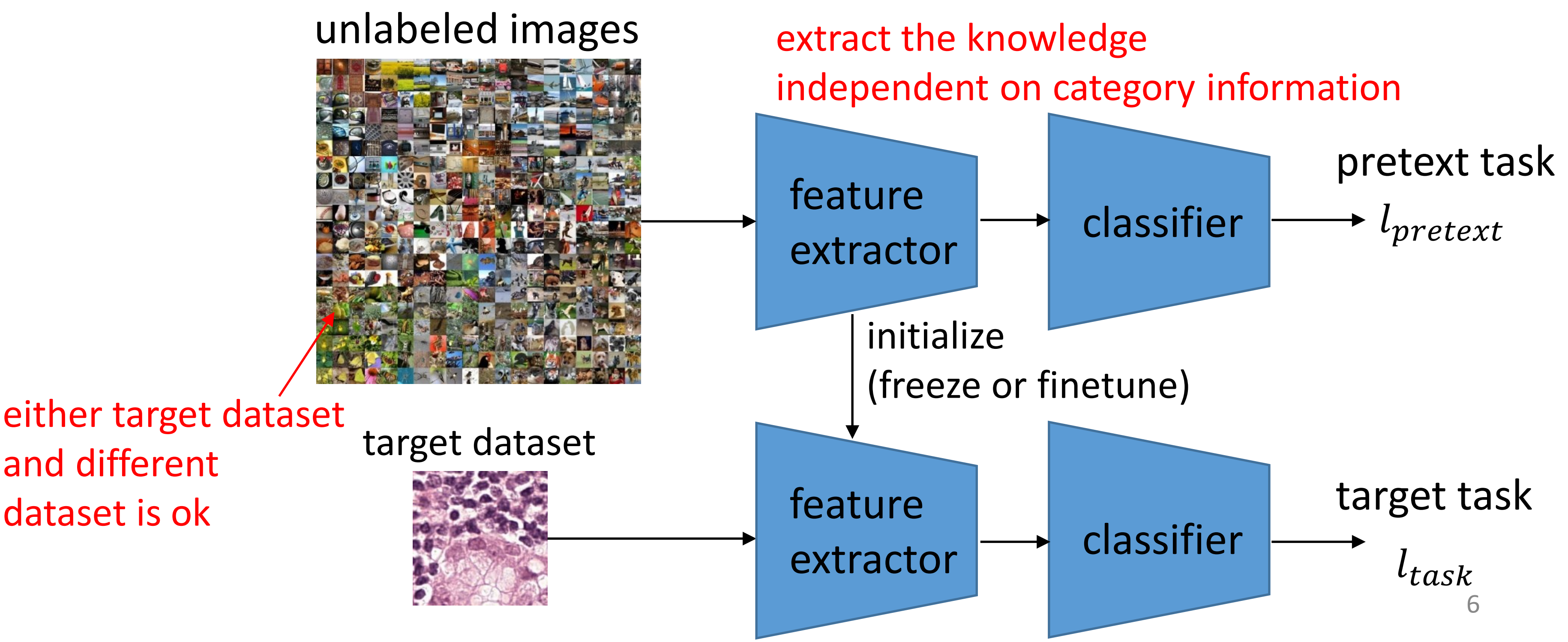


Today’s topic

- Self-supervised learning on equivariant neural networks
  https://arxiv.org/pdf/2303.04427.pdf
- Time series prediction considering hierarchical permutation equivariance https://arxiv.org/pdf/2305.08073.pdf

Self-Supervised Learning

- First we pretrain the model with user defined pretext task that does not use image label.
- Then we use the feature extractor for the target task.

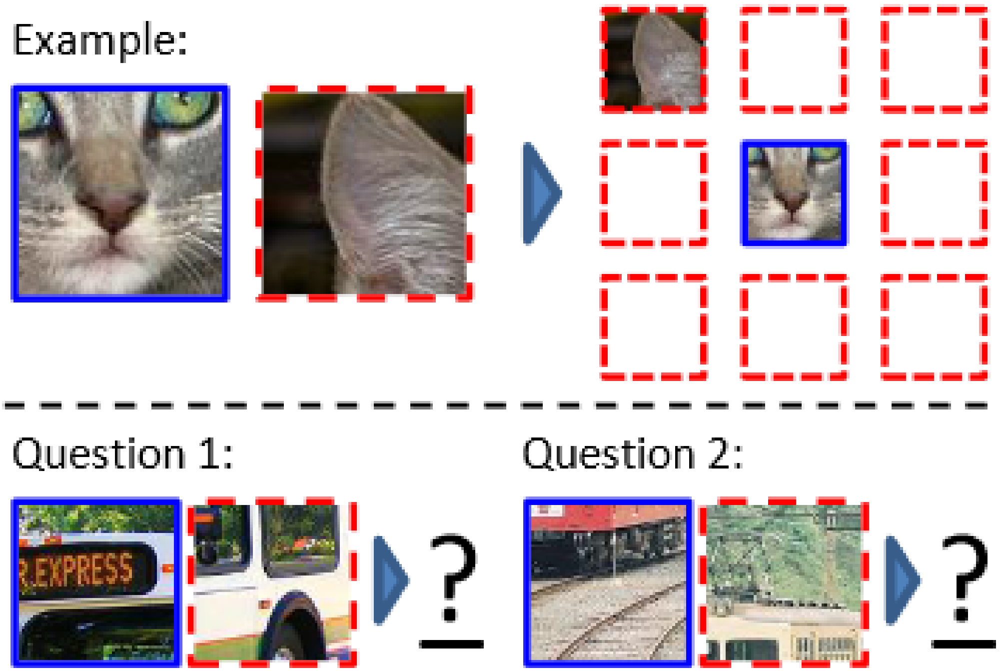


Self-Supervised Learning Methods

- Hand-crafted tasks
  - Train the model to solve hand-crafted ill-posed problem.
  - We assume that the feature extractor learn good image prior while trying to solve the problem.
- Contrastive learning
  - Apply data augmentation and trains the model to make the augmented images from the same image close.

Context prediction [Doersch et al., 2015]

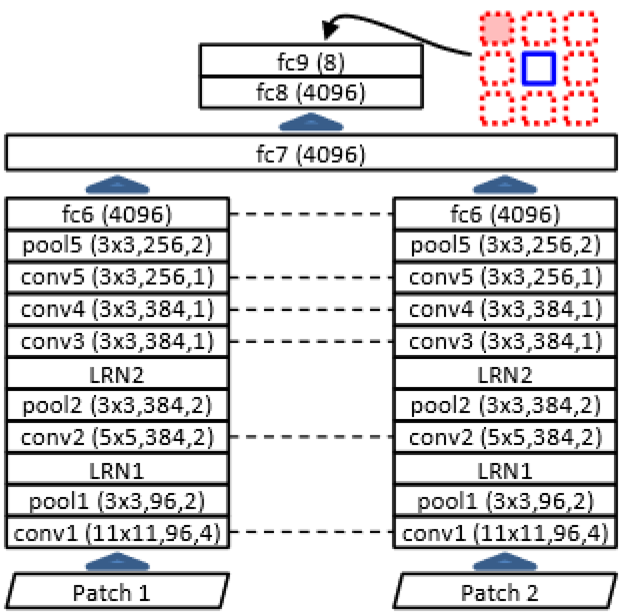
- Predict the spatial relationship between two image patches.

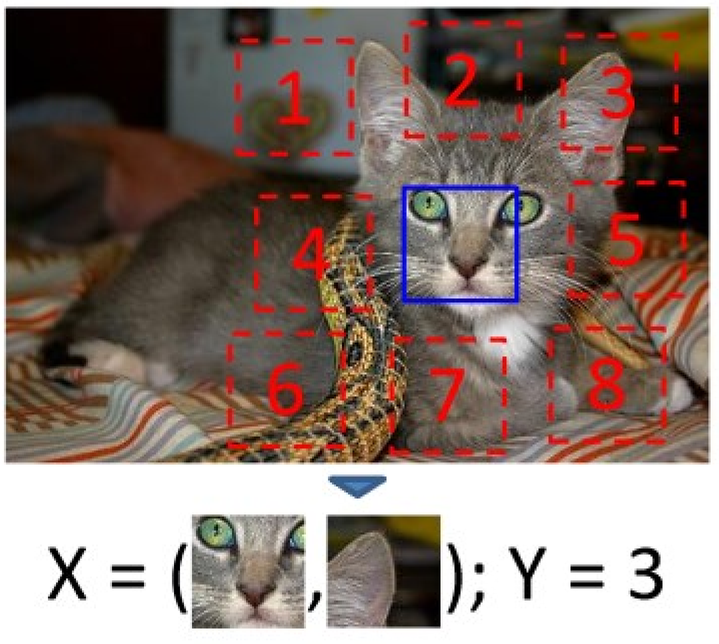


Model

- Cast the context prediction as 8 category prediction from two image sub-region.

architecture

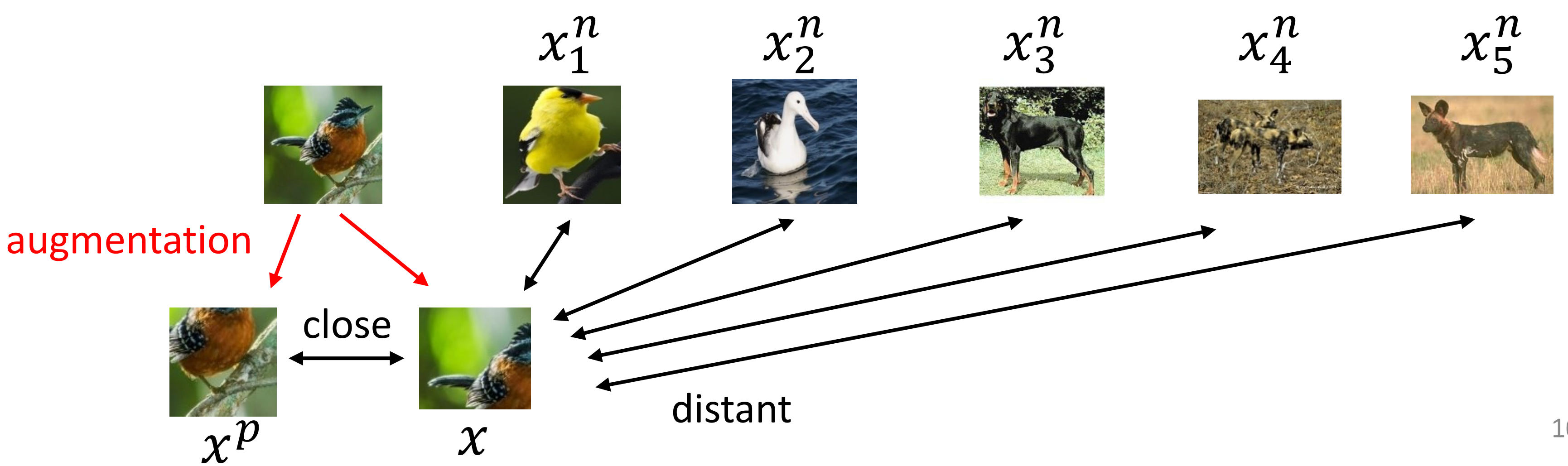




Contrastive Learning

- Learn feature so that two augmented image are closer than the other images by contrastive loss

$$\frac{\exp\left(\frac{x^t x^p}{\tau}\right)}{\exp\left(\frac{x^t x^p}{\tau}\right) + \sum_k \exp\left(\frac{x^t x_k^n}{\tau}\right)}$$



Motivation

- Combine the idea of
  - Exploiting the prior knowledge as group equivariant architecture.
  - Exploiting the prior knowledge as pretext task.

Difficulty

- The function learned by equivariant neural networks f_{NN} is restricted to equivariant such that $f_{NN}\left(T_{in}(x)\right) = T_{out}(f_{NN}(x))$.
- We cannot learn the task if the pretext label violates this equivariance.

input image x

transformed image $T(x)$

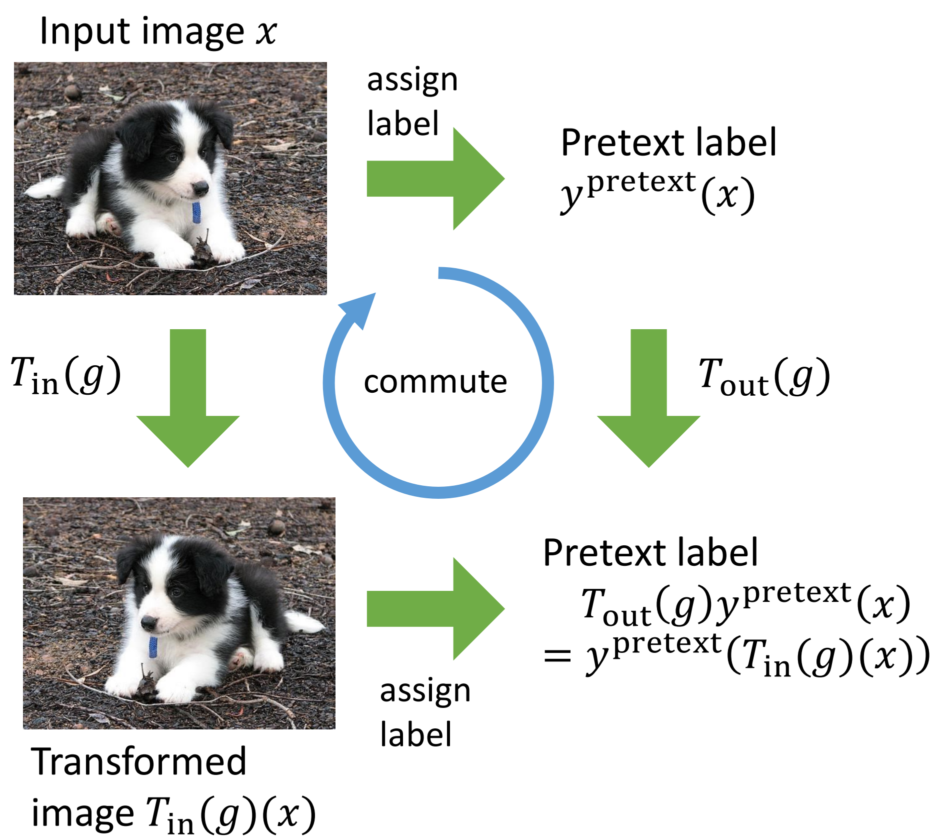
Needs to be consistent

Pretext label $y_x^{pretext}$ $\longleftrightarrow$ Pretext label $y_{T(x)}^{pretext}$

Proposed: equivariant pretext label

- Restrict the pretext label space so that satisfies

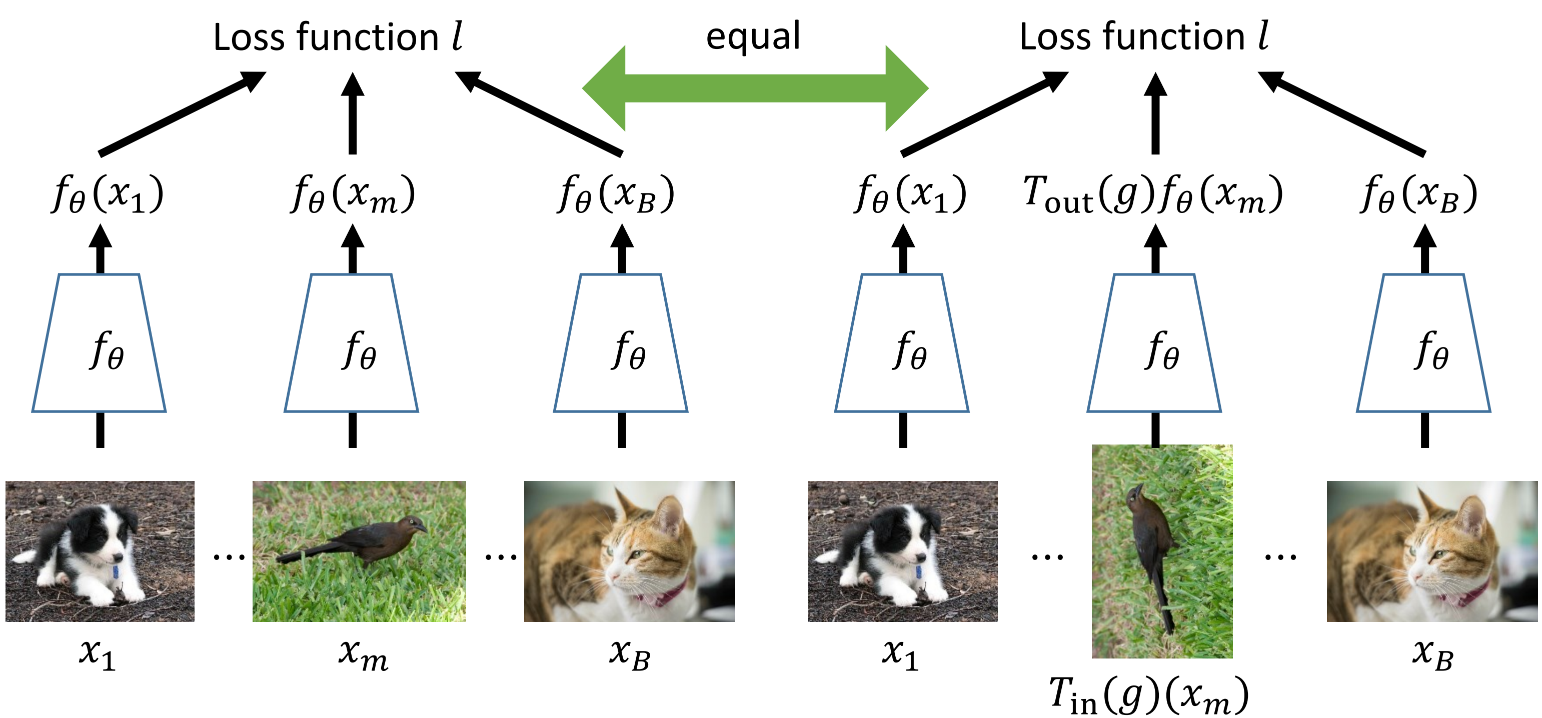
$$T_{\mathrm{out}}(g)y^{\mathrm{pretext}}(x) = y^{\mathrm{pretext}}(T_{\mathrm{in}}(g)(x)).$$



Proposed: invariant contrastive loss

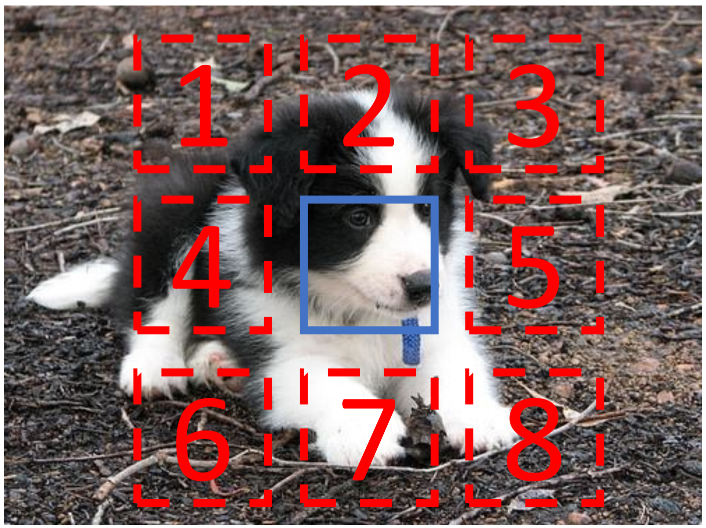
- Invariant contrastive loss is the loss function that satisfies

$$
\begin{aligned}
&l(f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_B)) \\
=&l(f_\theta(x_1), f_\theta(x_2), ..., T_{\mathrm{out}}(g) f_\theta(x_m), ..., f_\theta(x_B))
\end{aligned}
$$



Context prediction [Doersch et al., 2015]

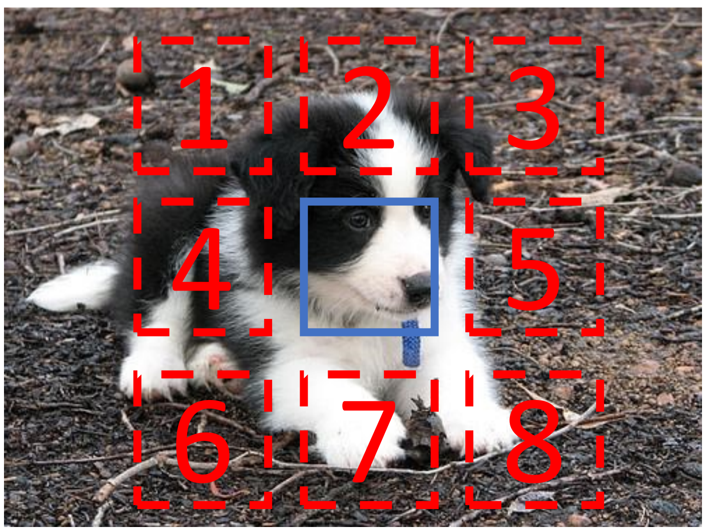
- Predict the spatial relationship between two image patches.



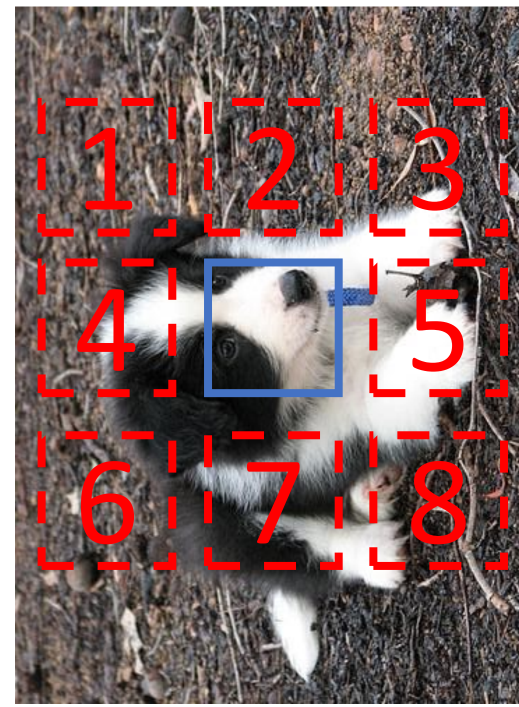
$X = ($  ,  $), Y = 4$

Equivariant Context Prediction

- The label space $\mathbb{R}^8$ of context prediction task satisfies 90° rotation equivariance.



90° rot



$X = ($  ,  $), Y = 4$

$$y_x^{pretext} = (0,0,0,1,0,0,0,0)$$

$X = ($  ,  $), Y = 7$

$$y_{T(x)}^{pretext} = (0,0,0,0,0,0,1,0)$$

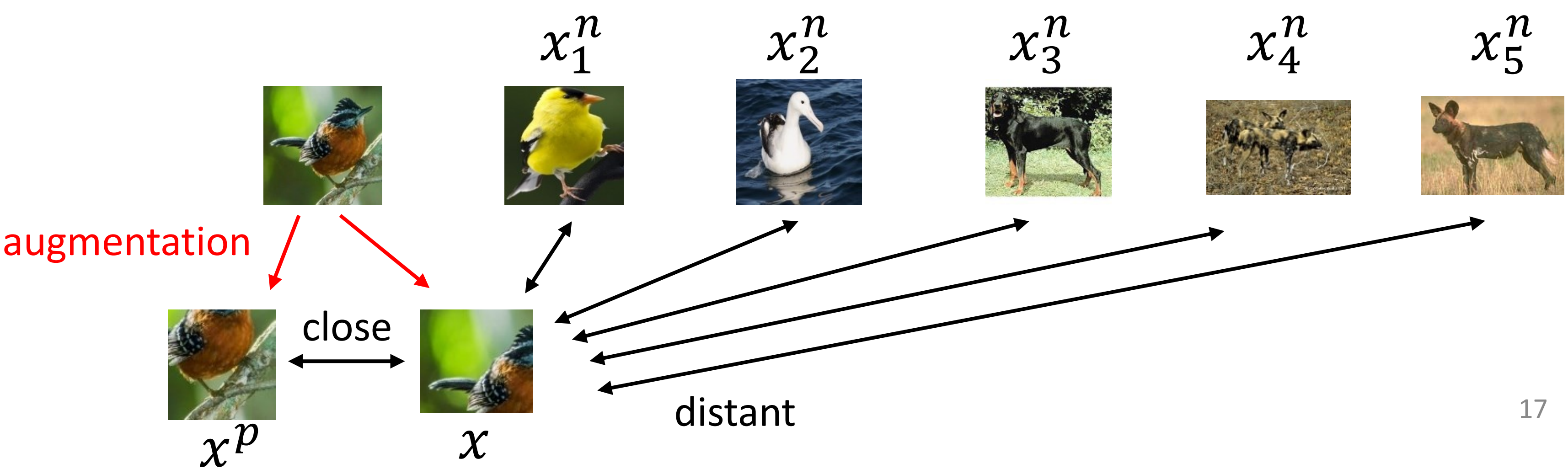
Invariant Contrastive Learning

- We average the output feature as

$$\frac{\exp\left(\frac{\left(\frac{1}{|G|}\sum_{g\in G} T_{out}(g)x\right)^t \frac{1}{|G|}\sum_{g\in G} T_{out}(g)x^p}{\tau}\right)}{\exp\left(\frac{\left(\frac{1}{|G|}\sum_{g\in G} T_{out}(g)x\right)^t \frac{1}{|G|}\sum_{g\in G} T_{out}(g)x^p}{\tau}\right)+\sum_k \exp\left(\frac{\left(\frac{1}{|G|}\sum_{g\in G} T_{out}(g)x\right)^t \frac{1}{|G|}\sum_{g\in G} T_{out}(g)x_k^n}{\tau}\right)}$$

to satisfy

$$\begin{aligned} &l(f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_B)) \\ =&l(f_\theta(x_1), f_\theta(x_2), ..., T_{\text{out}}(g) f_\theta(x_m), ..., f_\theta(x_B)) \end{aligned}$$



Experiment

- Evaluation:
  - Pretrain on ImageNet (1,300,000 images, 1,000 labels)
  - Apply linear classifier on top of the pretrained model.
- Architecture: ResNet50
- Compare
  - Standard non-equivariant model
  - Group equivariant model with the proposed loss
  - Group equivariant model with standard non-equivariant loss

Result

Table 1: Accuracy (%) on ImageNet with linear image classification setting.

Method	Baseline	Equivariant Model & Loss (Ours)	Equivariant Model Only
Context prediction	32.7	**35.1**	31.5
Jigsaw	35.1	**43.1**	42.5
Momentum Contrast	63.8	**65.7**	65.0
SwAV	71.4	**71.6**	68.2
SimSiam	65.9	**68.2**	65.5

Table 2: Mean AP (%) on VOC2007 with linear image classification setting.

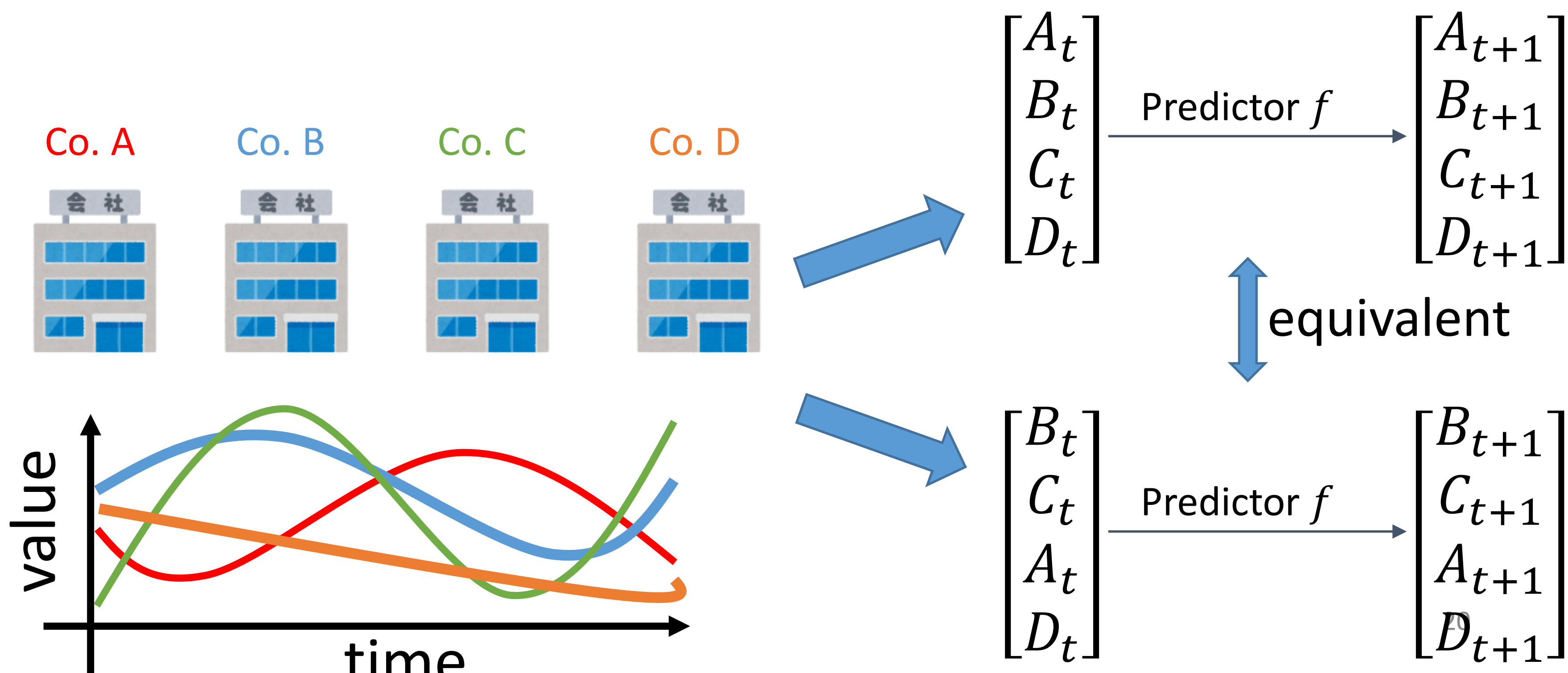
Method	Baseline	Equivariant Model & Loss (Ours)	Equivariant Model Only
Context prediction	51.7	**53.6**	49.1
Jigsaw	52.9	56.7	**57.8**
Momentum Contrast	80.7	**81.1**	80.2
SwAV	85.6	**86.8**	86.6
SimSiam	**81.7**	81.1	81.5

Table 3: Accuracy (%) on iNaturalist18 with linear image classification setting.

Method	Baseline	Equivariant Model & Loss (Ours)	Equivariant Model Only
Context prediction	**8.56**	**8.56**	6.97
Jigsaw	8.72	**13.8**	13.2
Momentum Contrast	33.4	**33.8**	31.2
SwAV	**42.1**	35.8	32.4
SimSiam	32.6	**33.7**	27.8

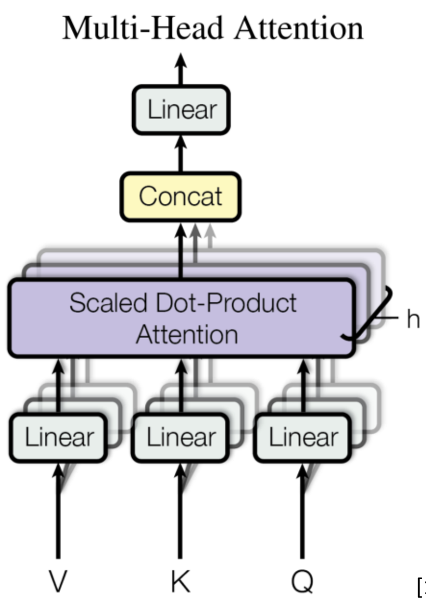
Multi variate time series prediction

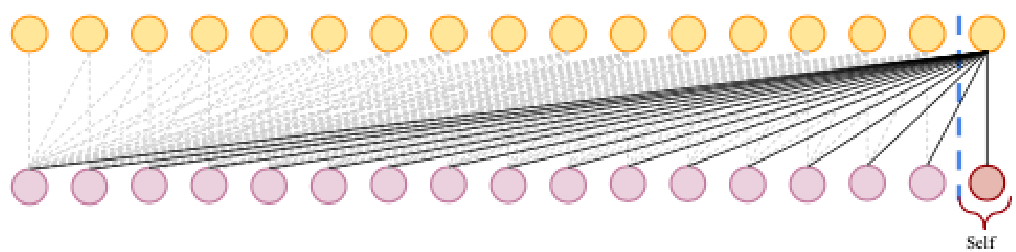
- Predict the future of multi-variate time series
- The predictor should be equivariant to the permutation of the time series.



Self-attention

- Calculate the output by the weighted average of the input, whose weights are calculated by the similarity of the inputs.
- We can preserve permutation equivariance by applying self-attention between the time series.



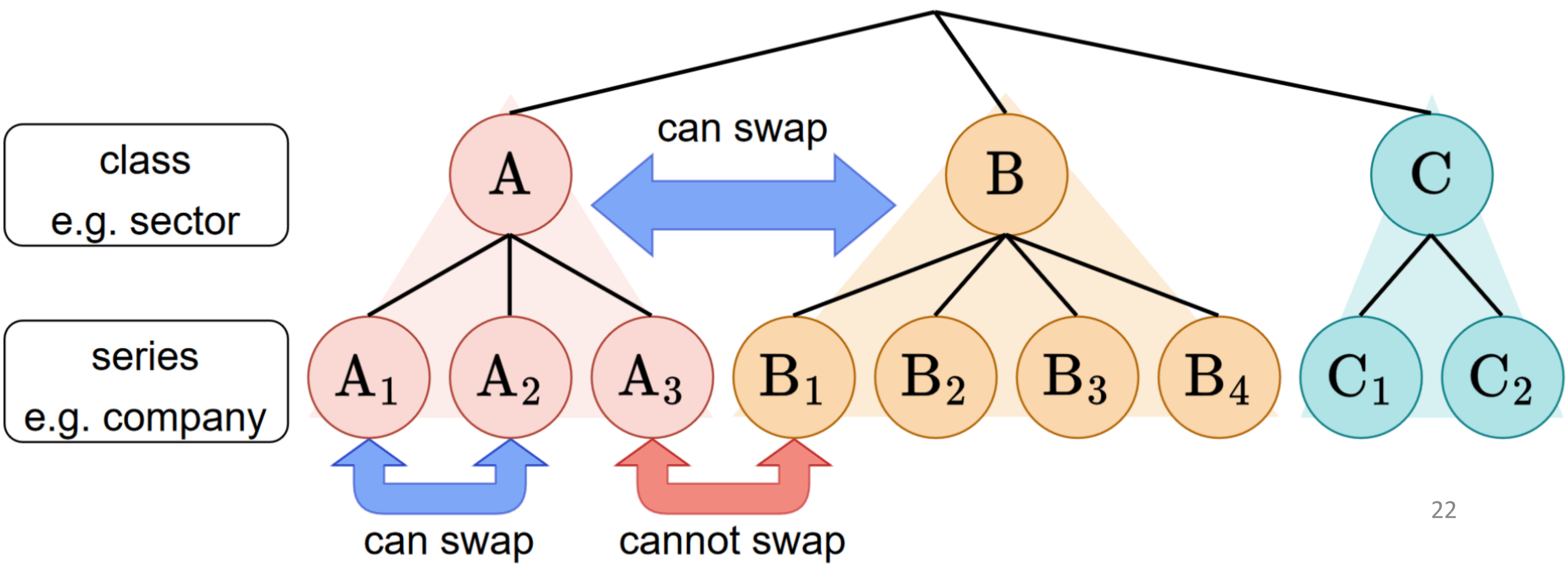


$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

[1] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems.* 2017.

Hierarchical permutation equivariance

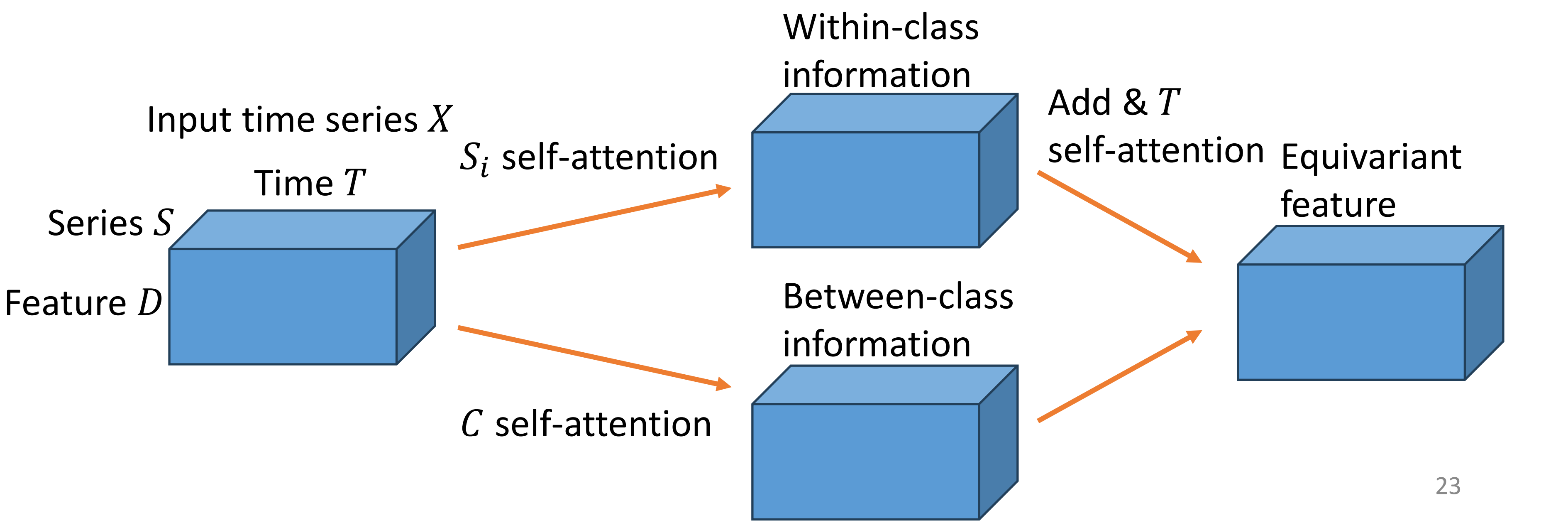
- We consider the case that time series are hierarchically grouped by such as sector, class.
- We want to restrict the equivariance to the permutation that considers hierarchy.



Feature extractor

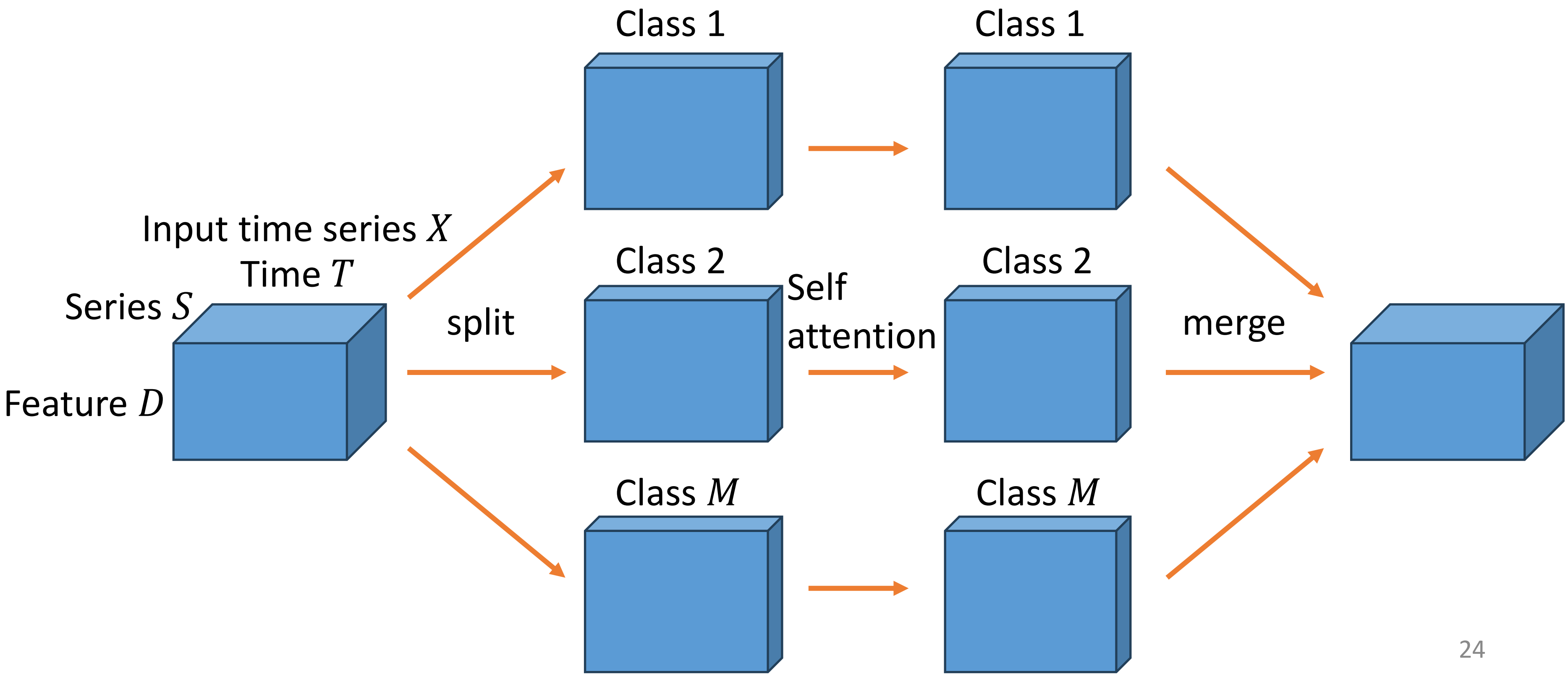
- Use 3D self-attention within class S_i, between classes C, time T.

$$\mathrm{SA3}(X) = \mathrm{SA_T}(\mathrm{SA_{S_i}}(X) + \mathrm{SA_C}(X))$$



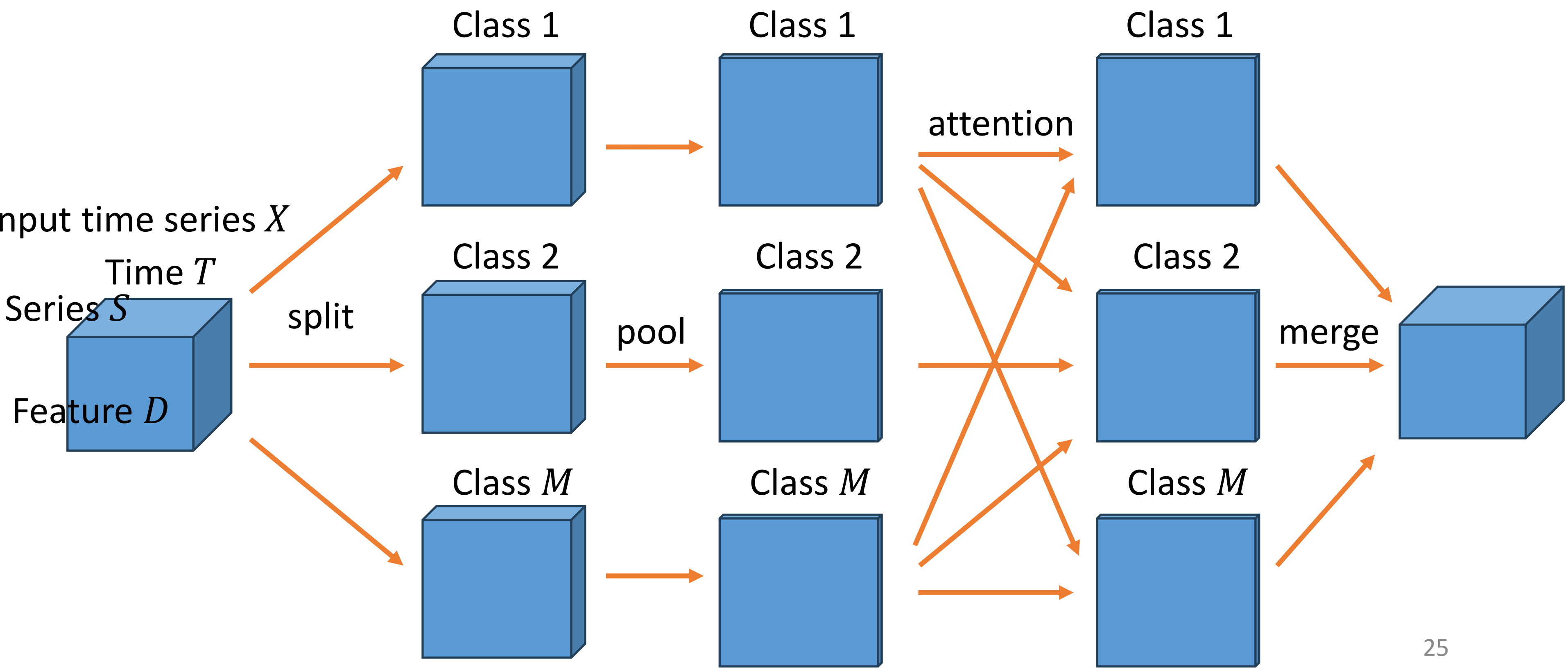
S_i self-attention

- Split the time series according to the class and apply self attention within class.



C self-attention

- Summarize the features of each class and then calculate the self-attention between the summarized features.



Experiment

- NBA: Trajectory of players and the ball in the basket game.
  - In: 40 steps, Out: 10 steps
  - 11 agents, 3 classes (ball, team A, team B)
  - Train: 80,000, Validation: 48,299, Test: 13,464
  - Evaluate the prediction accuracy by reducing the number of team A/B players at test time.

Result

ADE \ B A	0	1	2	3	4
0	(1.64±0.01 - 1.65±0.00)	1.69±0.01 - 1.69±0.01	1.73±0.01 - 1.73±0.01	1.78±0.01 - 1.78±0.01	1.85±0.01 - **1.84±0.01**
1	**1.67±0.01** - 1.68±0.01	**1.71±0.01** - 1.72±0.01	**1.76±0.01** - 1.77±0.01	1.83±0.01 - 1.83±0.01	1.92±0.01 - **1.91±0.01**
2	**1.69±0.01** - 1.71±0.01	**1.74±0.01** - 1.76±0.01	**1.81±0.01** - 1.82±0.01	1.89±0.01 - 1.89±0.01	2.01±0.01 - 2.01±0.01
3	**1.73±0.01** - 1.74±0.01	**1.79±0.01** - 1.80±0.01	**1.87±0.01** - 1.88±0.01	**1.98±0.01** - 1.99±0.01	**2.15±0.01** - 2.16±0.01
4	**1.76±0.01** - 1.79±0.01	**1.84±0.01** - 1.87±0.01	**1.95±0.01** - 1.97±0.01	**2.11±0.01** - 2.13±0.01	**2.38±0.01** - 2.40±0.01

FDE \ B A	0	1	2	3	4
0	(3.69±0.02 - 3.68±0.01)	3.78±0.02 - **3.76±0.01**	3.87±0.02 - **3.84±0.01**	3.98±0.02 - **3.94±0.01**	4.13±0.02 - **4.07±0.01**
1	3.73±0.02 - **3.72±0.01**	3.82±0.02 - **3.80±0.01**	3.93±0.02 - **3.90±0.01**	4.06±0.02 - **4.02±0.01**	4.25±0.02 - **4.19±0.01**
2	3.76±0.02 - 3.76±0.01	3.86±0.02 - **3.85±0.01**	3.99±0.02 - **3.97±0.01**	4.16±0.02 - **4.13±0.01**	4.41±0.02 - **4.35±0.01**
3	3.81±0.02 - 3.81±0.01	3.93±0.02 - 3.93±0.01	4.09±0.02 - **4.08±0.01**	4.31±0.02 - **4.29±0.01**	4.65±0.02 - **4.61±0.01**
4	**3.86±0.02** - 3.87±0.01	**4.01±0.02** - 4.02±0.01	4.22±0.02 - 4.22±0.02	4.53±0.02 - **4.52±0.02**	5.06±0.02 - **5.02±0.02**

NLL \ B A	0	1	2	3	4
0	(20.96±0.20 - 21.14±0.26)	21.43±0.29 - **21.40±0.23**	21.89±0.32 - **21.79±0.24**	22.45±0.35 - **22.26±0.24**	22.97±0.39 - **22.67±0.24**
1	**21.39±0.30** - 21.42±0.24	21.74±0.31 - **21.69±0.23**	22.29±0.35 - **22.16±0.24**	22.99±0.40 - **22.75±0.25**	23.69±0.45 - **23.31±0.25**
2	21.68±0.33 - 21.68±0.22	22.10±0.34 - **22.01±0.22**	22.78±0.39 - **22.60±0.23**	23.68±0.45 - **23.36±0.24**	24.65±0.52 - **24.15±0.23**
3	22.04±0.37 - **22.01±0.22**	22.57±0.39 - **22.44±0.21**	23.45±0.45 - **23.19±0.22**	24.66±0.54 - **24.22±0.23**	26.12±0.65 - **25.44±0.22**
4	22.51±0.41 - **22.43±0.20**	23.21±0.44 - **23.00±0.19**	24.39±0.53 - **24.02±0.19**	26.13±0.65 - **25.51±0.21**	28.57±0.84 - **27.56±0.17**

Left: without class information, Right: with class information
0~4 indicates the number of reduced players from each team

Conclusion

- We introduced two recent works relating equivariant neural networks.
    - Propose the idea of equivariant pretext labels and invariant contrastive loss to combine equivariant neural networks and self-supervised learning https://arxiv.org/pdf/2303.04427.pdf
    - Propose the multi-variate time series prediction method considers hierarchical permutation equivariance https://arxiv.org/pdf/2305.08073.pdf