Group Equivariant Stand-Alone Self-Attention for Vision

David W. Romero[1], Jean-Baptiste Cordonnier[2]

[1] VU Vrije Universiteit Amsterdam [2] EPFL

David W. Romero[1]

Jean-Baptiste Cordonnier[2]

1. VU VRIJE UNIVERSITEIT AMSTERDAM

2. EPFL

Group Equivariant Stand-Alone Self-Attention

David W. Romero[1], *Jean-Baptiste Cordonnier*[2]

Motivation: Self-attention has been successful in several applications. However, it does not leverage known symmetries of the data modality to enhance its sample-efficiency.
How can we induce group equivariance to self-attention networks, e.g., Transformers?

Results:
Theoretical:
1. We show that **only** the positional encoding must be modified.
2. Group self-attention is more expressive than group convolutions.
3. Global group self-attention is an equivariant universal approximator.
4. Group self-attention networks (GSA-Nets) are steerable by nature.
5. We generalize the proposition that group equivariance is only obtained via group convolutions to a broader family of linear functions.

Empirical:
1. Consistent performance improvements of GSA-Nets over non-equivariant ones**.**
2. GSA-Nets converge much faster than non-equivariant ones.
3. G-CNNs still perform better than GSA-Nets (for now 😉)
4. Time and memory complexity of group self-attention is a very restrictive bottleneck.

1. VU VRIJE UNIVERSITEIT AMSTERDAM 2. EPFL

Group Equivariant Stand-Alone Self-Attention

Outline

- Preliminaries
  - Symmetries and equivariance
  - The Self-Attention Operation

- Group Equivariant Stand-Alone Self-Attention
  - How can we impose group equivariance to self-attention?
  - Lifting & Group self-attention
  - Theoretical Results

- Experimental Results

- Conclusions

Group Equivariant Stand-Alone Self-Attention

Symmetries and Equivariance

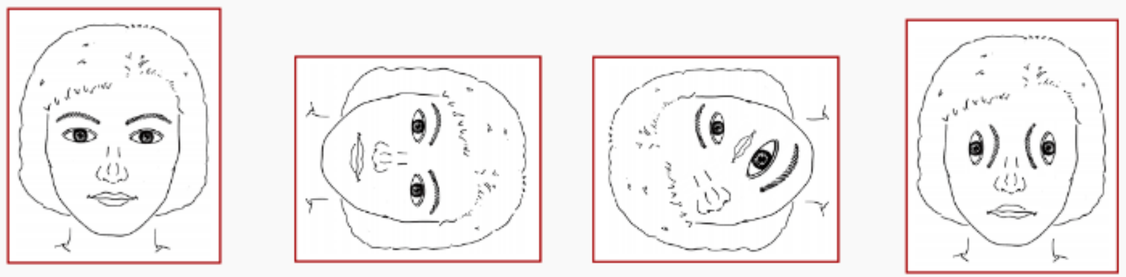
A **symmetry group** is a set of transformations that leave an object invariant.

- Translations
- Rotations
- Scaling
- Change of pitch in audio



Invariance: If a function ***f*** is invariant to the transformations, it **does not** retain additional information other than presence/absence of a pattern: ***{cat, cat, cat}***

Equivariance: If ***f*** is equivariant, it **does** retain information about transformations applied to the input: ***{cat 1x 0°, cat 0.5x 30°, cat 1x -45°}***

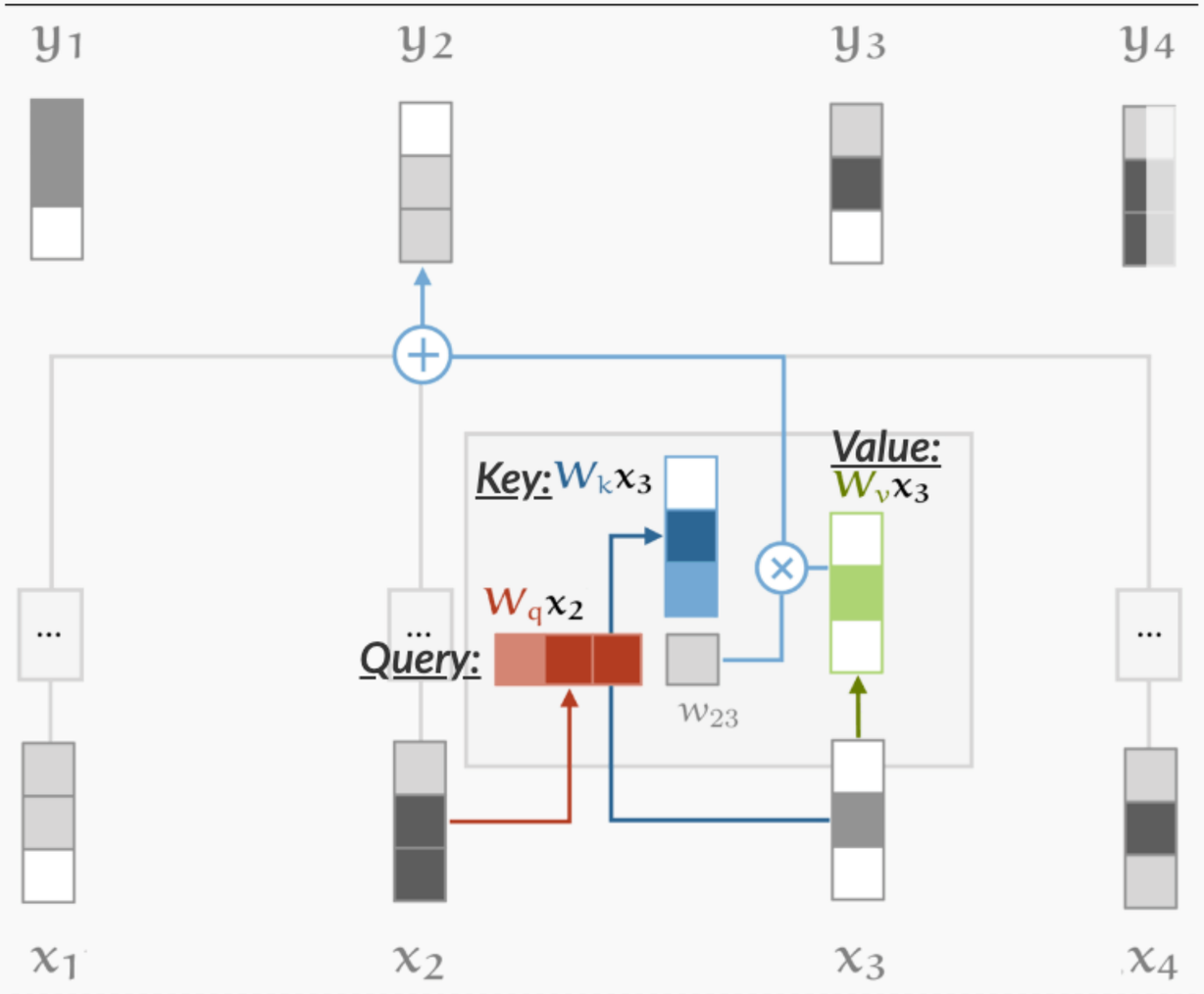


Group Equivariant Stand-Alone Self-Attention

Self-Attention Formulation

1. Self-attention takes a query token and compares it to all other key tokens in the sequence (query-key scores).

2. Based on the query-key scores, the value of the query token is updated by the value of all other key tokens.



Parameters:

W_q W_k W_v

Not tied to any position!

Question:

Where is position encoded?

Replace

$$w_{23} = W_q x_2 (W_k x_3)^T$$

by

$$w_{23} = W_q x_2 (W_k (x_3 + \rho(2,3))^T$$

where

$\rho :=$ ***positional encoding***

Group Equivariant Stand-Alone Self-Attention

How can we induce group equivariance to self-attention?

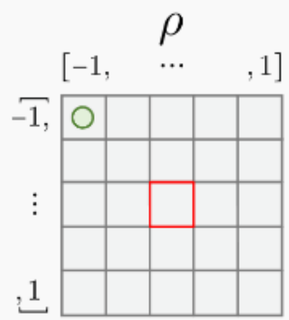
Main Observation.
The only "geometric" part of self-attention is the positional encoding.
Hence, this is the only part that needs to be modified.

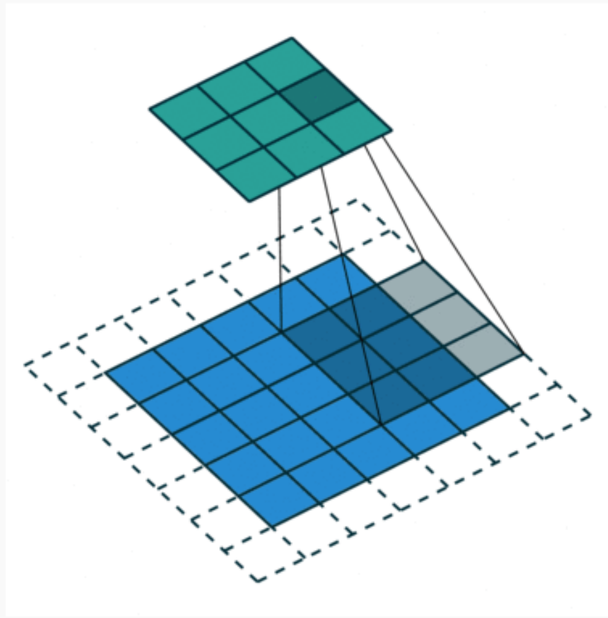
1. The positional encoding must be invariant to the action of the group.

Examples:

1. ***Relative positional encoding*** is invariant to translations.
2. ***Constant positional encoding*** is invariant to permutations.

Positional Encoding:





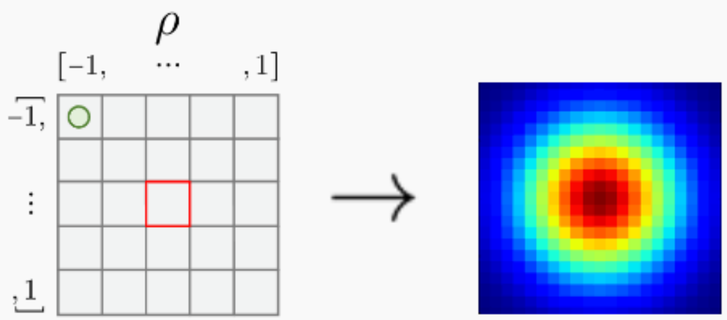
Group Equivariant Stand-Alone Self-Attention

How can we induce group equivariance to self-attention?

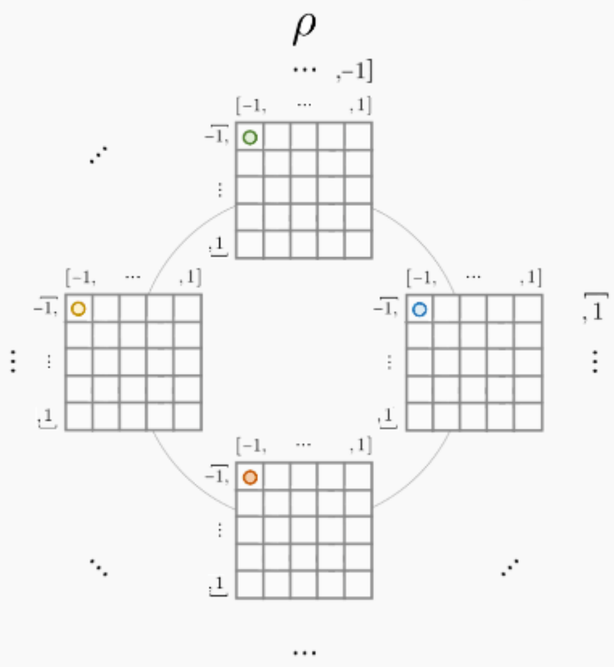
Main Observation.
The only "geometric" part of self-attention is the positional encoding.
Hence, this is the only part that needs to be modified.

2. For maximum expressiveness, the positional encoding must be defined on the group itself.

Positional Encoding:



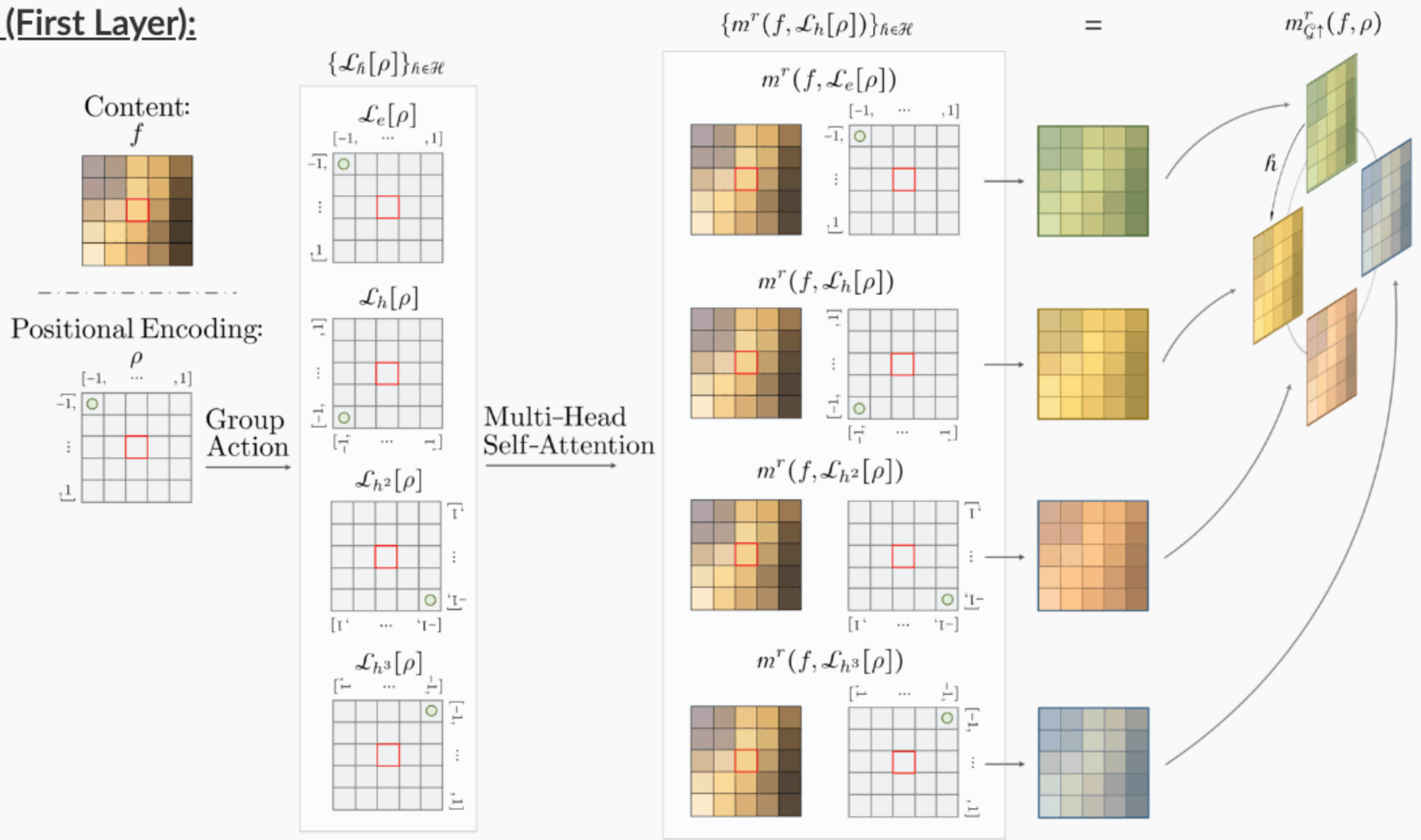
Positional Encoding:



Group Equivariant Stand-Alone Self-Attention

David W. Romero[1], *Jean-Baptiste Cordonnier*[2]

Method (First Layer):



1. Modify the positional encoding by the action of each element of the group.

2. Compute self-attention with each of these positional encodings.

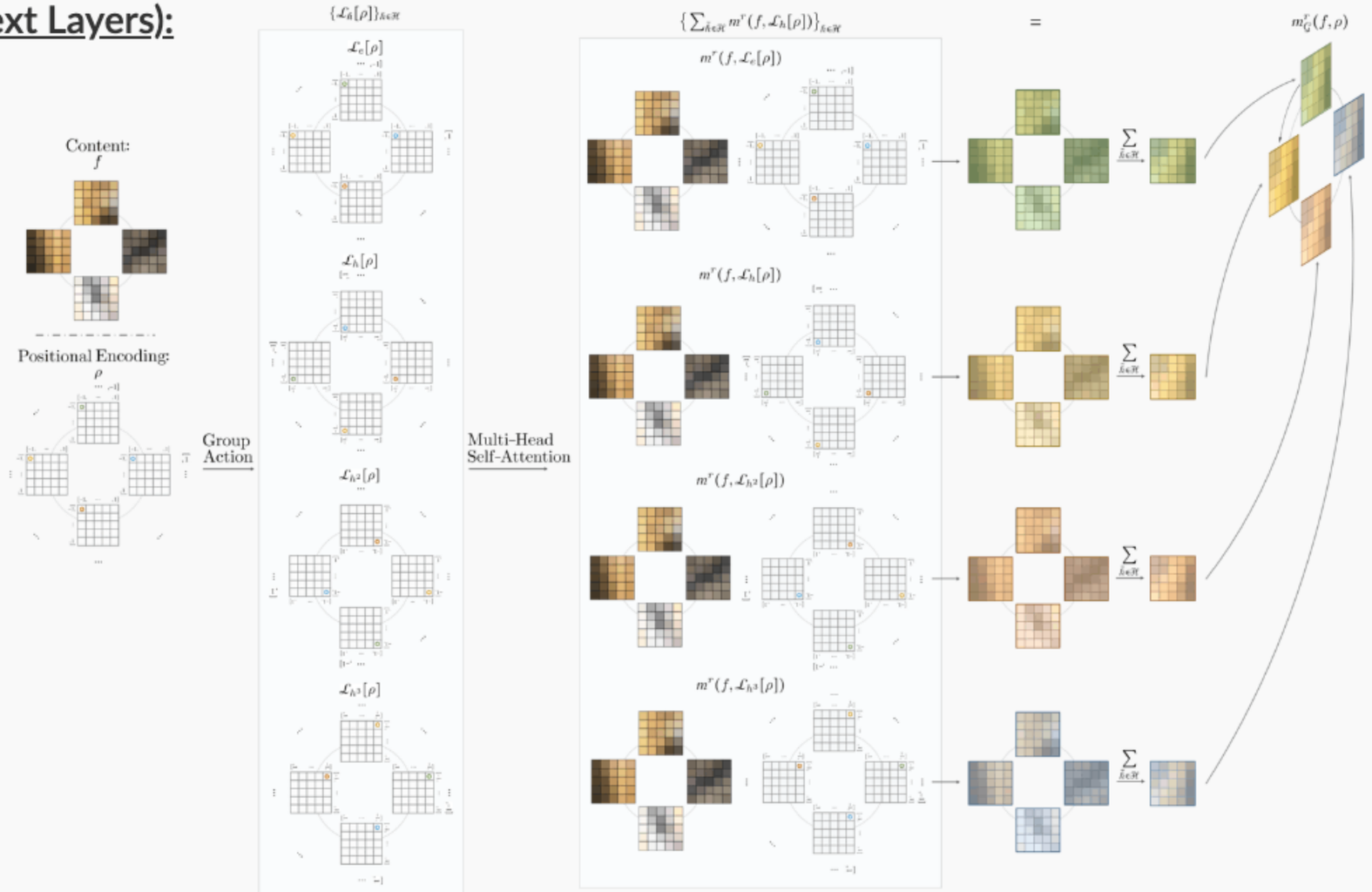
3. Concatenate the responses

1. VU VRIJE UNIVERSITEIT AMSTERDAM 2. EPFL

Group Equivariant Stand-Alone Self-Attention

David W. Romero[1], *Jean-Baptiste Cordonnier*[2]

Method (Next Layers):



1. Modify the positional encoding by the action of each element of the group**.**

2. Compute self-attention with each of these positional encodings.

3. Concatenate the responses

1. VU VRIJE UNIVERSITEIT AMSTERDAM 2. EPFL

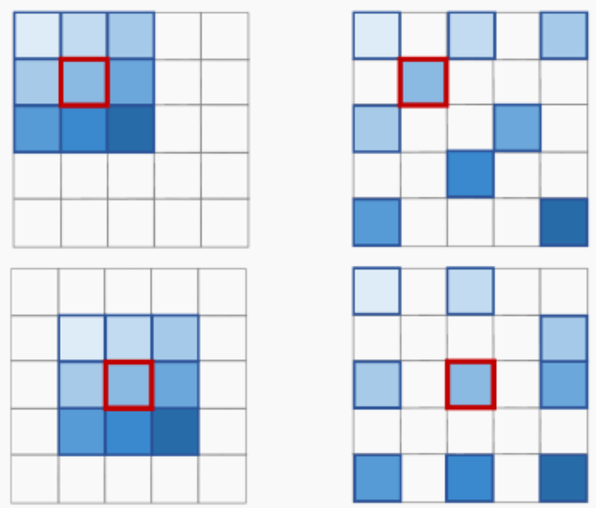
Group Equivariant Stand-Alone Self-Attention

Theoretical Results

1. Group Self-Attention is more expressive than group convolutions.

Cordonnier et al. (2020): *"Any convolutional layer can be described as a multi-head self-attention layer provided enough heads."*

- An analogy can be drawn for group variants of self-attention and convolution: *Given enough heads, group self-attention can describe any group convolutional filter.*
- **However**, since group self-attention typically handles larger receptive fields, the family of functions that can be described is also larger.



Group Equivariant Stand-Alone Self-Attention

Theoretical Results

2. Global self-attention is an equivariant universal approximator.

Ravanbakhsh (2020): *"Functions induced by regular group representations and global receptive fields are equivariant universal approximators."*

- Global group self-attention fulfills these conditions.

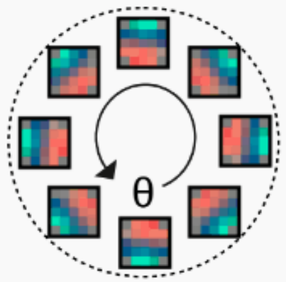
Group Equivariant Stand-Alone Self-Attention

Theoretical Results

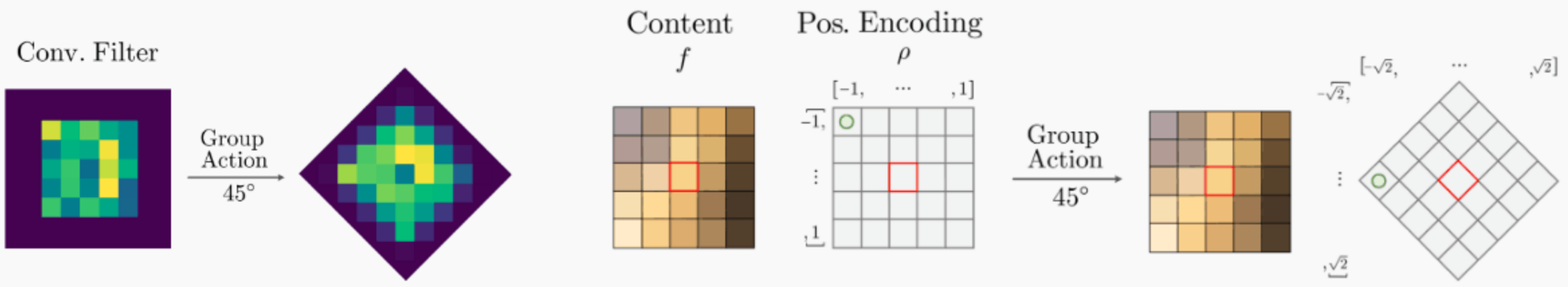
3. Group self-attention is steerable by nature.

Steerability: Several groups are not defined on a discrete grid, e.g., 45° rotations. Discrete kernels must be interpolated, which (strongly) deforms the filters.

Steerable networks, define kernels on a continuous basis to **avoid such artifacts**. Sampling instead of interpolating.



In group self-attention, groups act on the positional encoding. Since it is defined on a continuous space, **no interpolation is need**. Hence, **group self-attention is steerable**.



Group Equivariant Stand-Alone Self-Attention

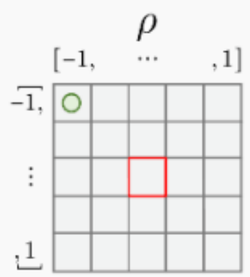
Theoretical Results

4. Generalizing Kondor & Trivedi (2018)'s main statement:
"The group convolution is the only linear group equivariant map".

Based on our findings, we are able to generalize this statement as:
"Linear mappings on G, whose positional encoding is G-invariant are G-equivariant"

Examples:

1. Self-attention with relative positional encodings (*translation equivariance*).
2. Self-attention with constant positional encoding (*permutation equivariance*).
3. Convolution (*translation equivariance*) -Look at the filter parametrization-.
4. Group convolution (*group equivariance*) -Look at the filter parametrization-.



$$[f \star_{\mathbb{R}^d} \psi](y) = \int_{\mathbb{R}^d} f(x)\psi(x-y)\,\mathrm{d}x \qquad [f \star_{\mathcal{G}} \psi](g) = \int_{\mathcal{G}} f(\tilde{g})\psi(g^{-1}\tilde{g})\,\mathrm{d}\mu(\tilde{g})$$

Attentive Group Equivariant Convolutional Networks

Experimental Results

Table 2: Classification results. All convolutional architectures use `3x3` filters.

rotMNIST

Model	Acc. (%)	Params.
Z2_SA	96.37	44.67k
R4_SA	97.46	
R8_SA	97.90	
R12_SA	**97.97**	
R16_SA	97.66	
Z2_CNN[+]	94.97	21.75k
R4_CNN[†]	98.21	77.54k
α-R4_CNN[†]	**98.31**	73.13k

[+]Cohen & Welling (2016)
[†]Romero et al. (2020a)

CIFAR10

Model	Acc. (%)	Params.
Z2_SA	82.3	2.99M
Z2M_SA	**83.72**	
Z2_CNN[+]	**90.56**	1.37M

[+]Cohen & Welling (2016).

PatchCamelyon

Model	Acc. (%)	Params.
Z2_SA	83.04	205.66k
R4_SA	83.44	
R8_SA	83.58	
R4M_SA	**84.76**	
Z2_CNN[†]	84.07	130.60k
R4_CNN[†]	87.55	129.65k
R4M_CNN[†]	88.36	124.21k
α_F-R4_CNN[†]	88.66	140.45k
α_F-R4M_CNN[†]	**89.12**	141.22k

[†]Romero et al. (2020a).

GSA-Nets perform consistently better than vanilla self-attention networks.

However, G-CNNs still outperform GSA-Nets.

We conjecture that this is due to the more difficult optimization problem in GSA-Nets (optimization over positions to attend to).

Attentive Group Equivariant Convolutional Networks

Experimental Results

Table 1: Accuracy vs. neighbourhood size.

	ROTMNIST		
MODEL	NBHD. SIZE	ACC. (%)	TRAIN. TIME / EPOCH
R4_SA	3x3	96.56	04:53 - 1GPU
	5x5	**97.49**	05:34 - 1GPU
	7x7	97.33	09:03 - 1GPU
	9x9	97.42	09:16 - 1GPU
	11x11	97.17	12:09 - 1GPU
	15x15	96.89	10:27 - 2GPU
	19x19	96.86	14:27 - 2GPU
	23x23	97.05	06:13 - 3GPU
	28x28	96.81	12:12 - 4GPU

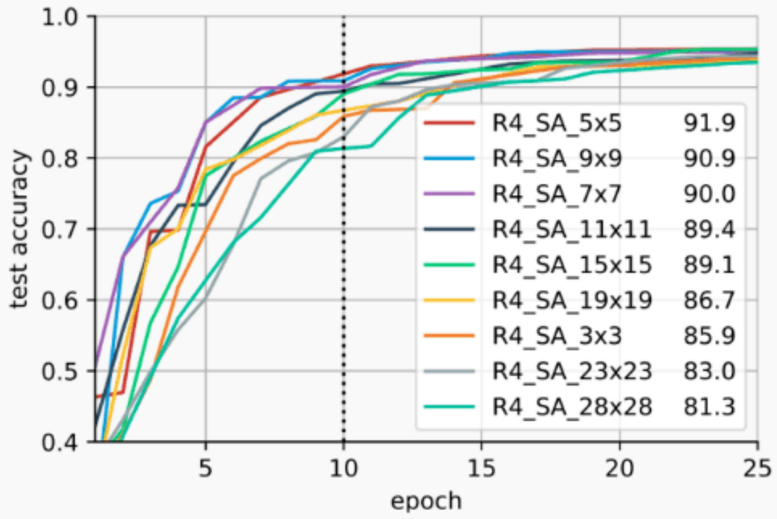


Figure 2: Test accuracy in early training stage.

1. The optimization problem seems, indeed, to become more complex for larger neighborhood sizes.

2. Time and memory complexity of group self-attention is a very restrictive bottleneck.

Group Equivariant Stand-Alone Self-Attention

Conclusions

Based on our theoretical results, we believe that GSA-Nets have the potential to replace G-CNNs and become the standard solution for group equivariant applications. **This also holds for SA-Nets and conventional CNNs.**

Though theoretical results are promising, further research in terms of architectural design, optimization strategies and stability is required. **This holds for self-attention in general.**

Thank you!

David W. Romero[1], Jean-Baptiste Cordonnier[2]

VU VRIJE UNIVERSITEIT AMSTERDAM

EPFL