ETH zürich

Attention as a hypernetwork

Simon Schug *

Seijin Kobayashi

Yassir Akram

João Sacramento **

Razvan Pascanu **

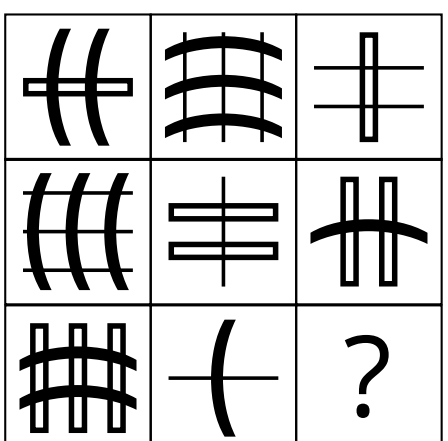
* *sschug@ethz.ch* ** *shared senior authors*

Overview

How do transfomers compositionally generalize?

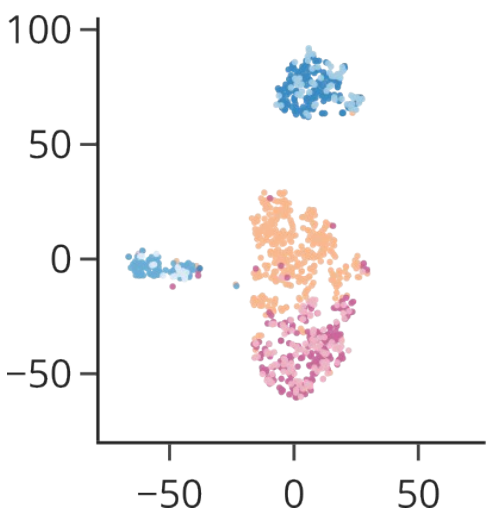
Part I: Symbolic Raven

An abstract reasoning task to study compositional generalization



Part II: Attention as a hypernetwork

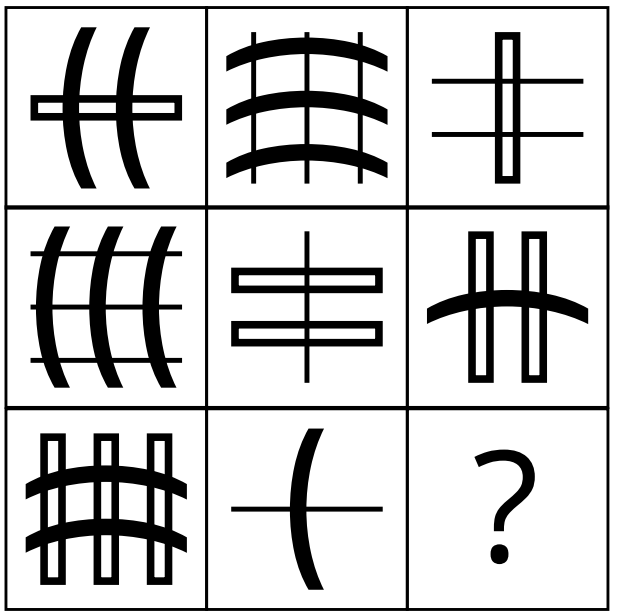
Treating attention as a hypernetwork reveals reusable & interpretable operations.



Raven's progressive matrices

A non-verbal test to measure human intelligence and abstract reasoning.[1,2]

- 3x3 matrices with the final panel hidden
- Require discovering one or several rules that govern the features encoded in each image
  - Simpler tasks can often be solved quickly through pattern recognition
  - More difficult tasks require generating and testing hypotheses
- Difficulty of *finding correspondences*[2]:
  - Which figural elements correspond to each other / are operated by the same rule

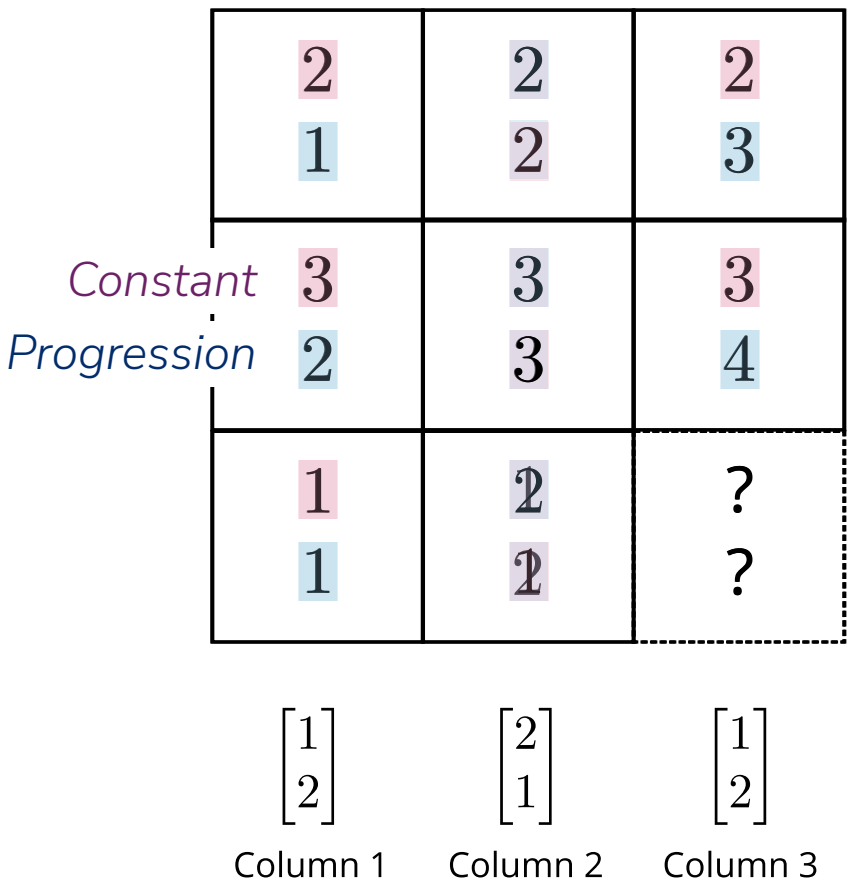


[1]Raven, 1936
[2]Carpenter et al., Psych. Rev., 1990

Symbolic Raven

Generating symbolically encoded Raven's matrices

- Each panel has *K* different features.
- The features of each panel change according to *K* rules within each row.
- Each rule is a function that takes two integers as input and outputs an integer.
- The features of each panel are permuted according to a random permutation which is fixed per column.



Symbolic Raven

Testing compositional generalization

We hold-out certain rule combinations from the training distribution in order to test if models can compositionally generalize.

Split rule combinations

Rules

A constant
B progression (+1)
C progression (+2)
D progression (-1)
E progression (-2)
F sum
G difference
H distribute-three

Splits

Train	AA	
	AB	
	AD	
	⋮	
75%	GH	

OOD	AC	
	⋮	
25%	HH	

1. Sample task rules

Combination of M rules

constant
progression (+1)

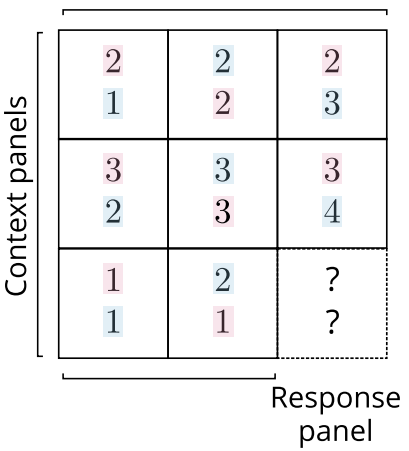
(e.g. **AB**)

Permutation

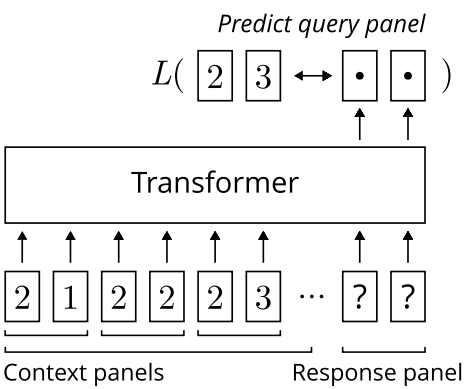
Each column permutes features to model the problem of finding correspondences.

$\begin{bmatrix}1\\2\end{bmatrix}$ Column 1 $\begin{bmatrix}2\\1\end{bmatrix}$ Column 2 $\begin{bmatrix}1\\2\end{bmatrix}$ Column 3

2. Sample task instance



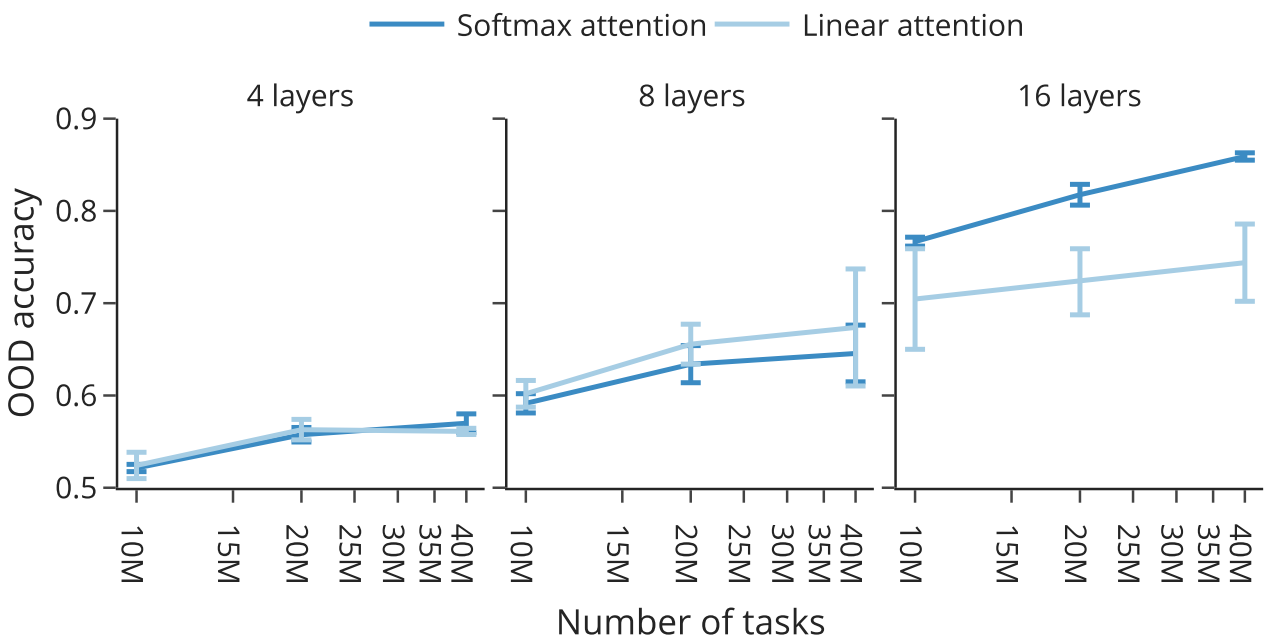
3. In-context learning



In-context learning

Scaling data and network size

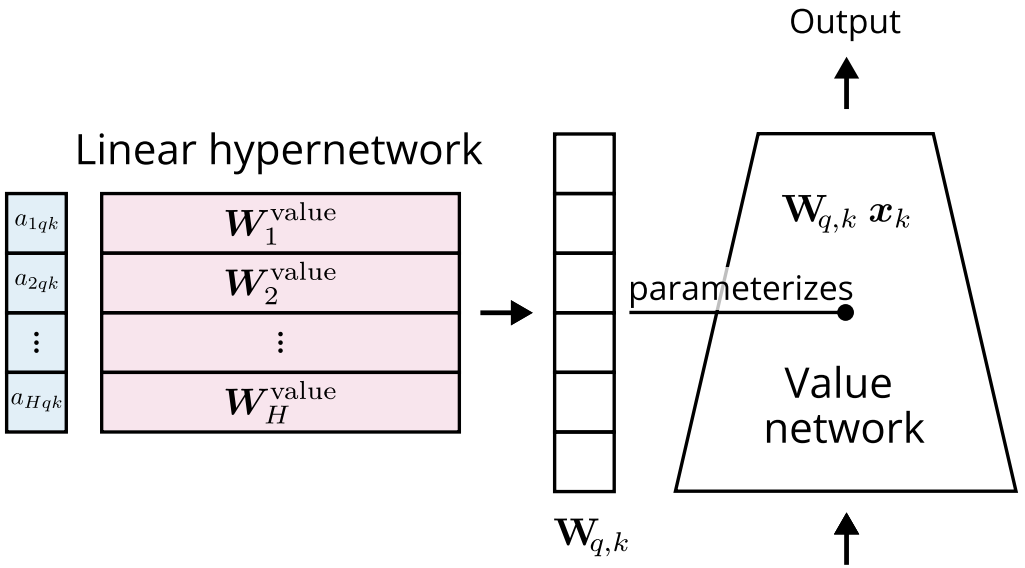
At sufficient scale, transformers can compositionally generalize on the Symbolic Raven task.

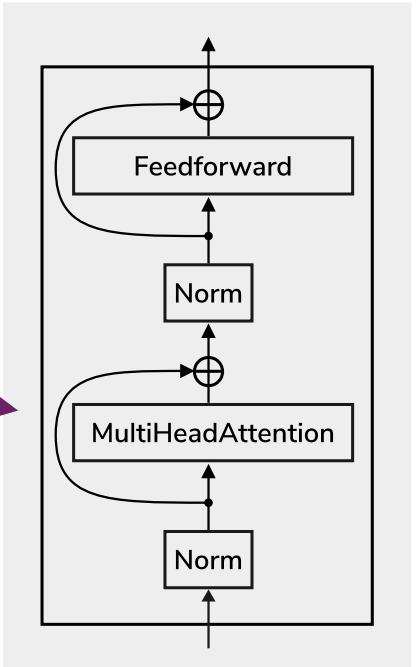


How do they do it?

Attention as a hypernetwork

Hypothesis: An implicit hypernetwork mechanism inside of multi-head attention supports composing reusable operations.



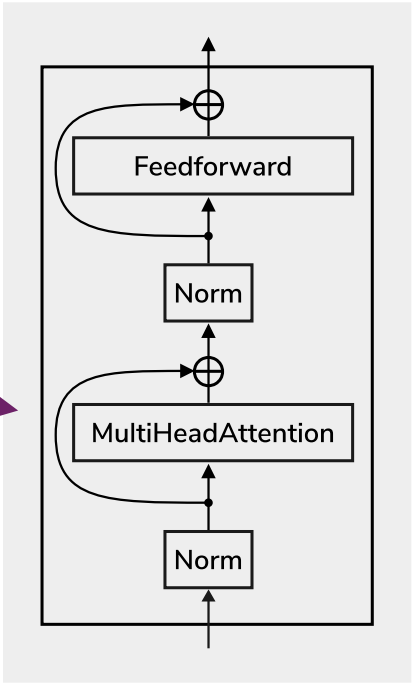


Transformer block

Attention as a hypernetwork

Hypothesis: An implicit hypernetwork mechanism inside of multi-head attention supports composing reusable operations.

$$
\begin{aligned}
\mathrm{MHA}_q(\mathbf{X}) :=& \boldsymbol{W}^{\text{out}} \bigoplus_{h=1}^{H} \sum_{k=1}^{T} a_{h,q,k}\ \boldsymbol{W}_h^{\text{value}} \mathbf{x}_k \\
&= \sum_{h=1}^{H} \boldsymbol{W}_h^{\text{out}} \sum_{k=1}^{T} a_{h,q,k}\ \boldsymbol{W}_h^{\text{value}} \mathbf{x}_k \\
&= \sum_{k=1}^{T} \left(\sum_{h=1}^{H} \underbrace{a_{h,q,k}}_{\text{latent code}}\ \underbrace{\boldsymbol{W}_h^{\text{out}} \boldsymbol{W}_h^{\text{value}}}_{\text{hypernetwork}} \right) \mathbf{x}_k \\
&= \sum_{k=1}^{T} \underbrace{\mathbf{W}_{q,k}}_{\text{value network}}\ \mathbf{x}_k
\end{aligned}
$$

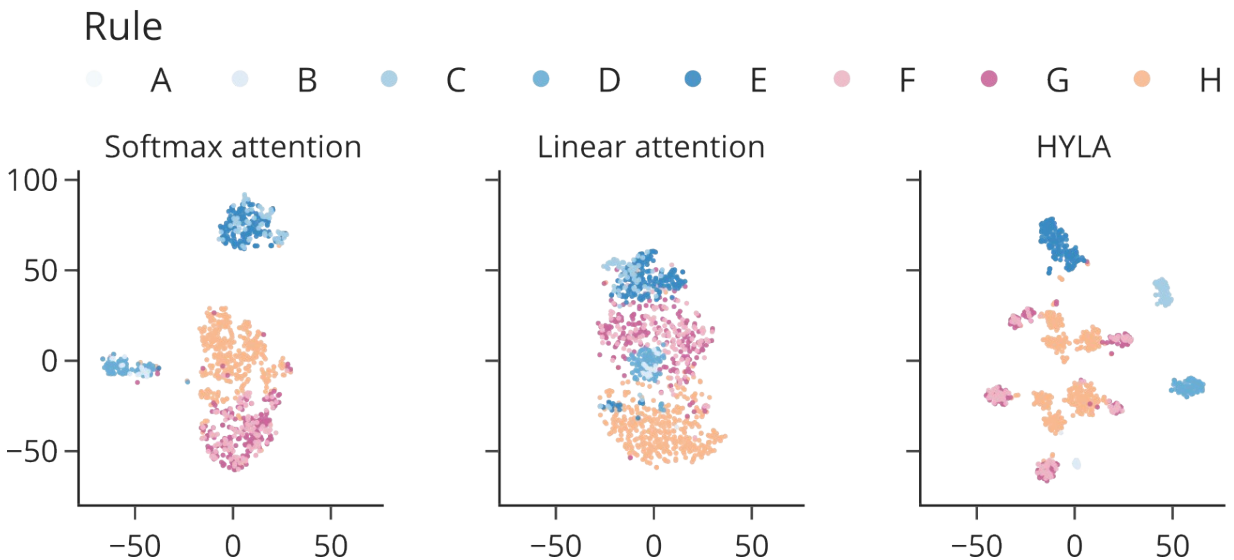


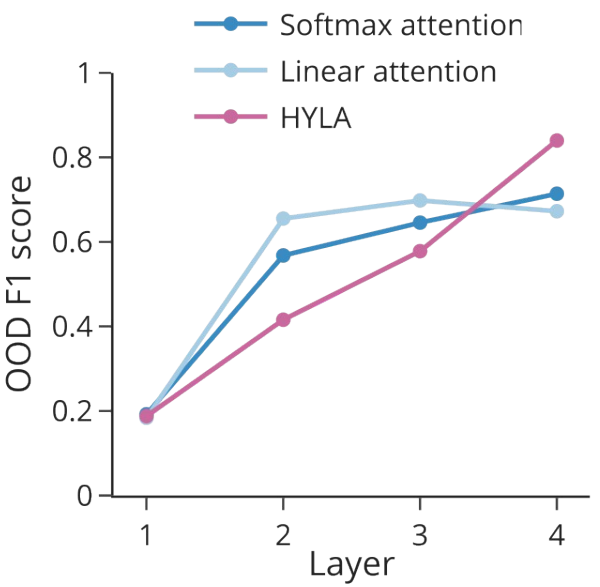
Transformer block

Attention as a hypernetwork

Latent code structure

The latent code is functionally structured and predictive of the task rules.





HYpernetwork Linear Attention

Reinforcing the hypernetwork mechanism

The value network of standard multi-head attention is linear.
We can make it nonlinear without introducing additional parameters.

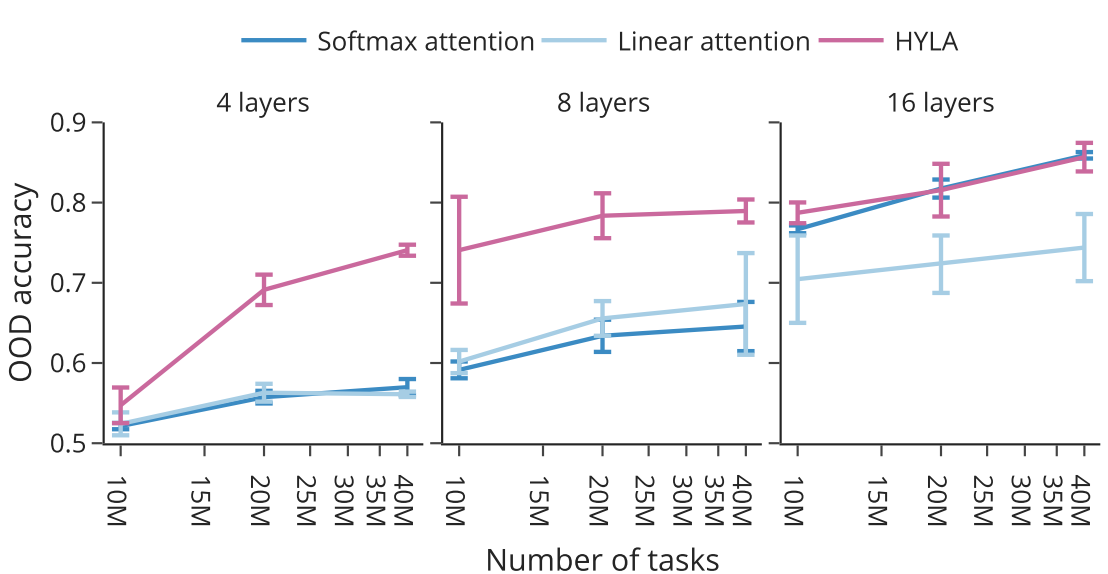
$$\begin{aligned}\mathrm{HYLA}_q(\mathbf{X}) &= \sum_{k=1}^{T} \left(\sum_{h=1}^{H} a_{h,q,k} \; \boldsymbol{W}_h^{\text{out}} \right) \phi \left(\sum_{h=1}^{H} a_{h,q,k} \; \boldsymbol{W}_h^{\text{value}} \mathbf{x}_k \right) \\ &= \sum_{k=1}^{T} \mathbf{W}'_{q,k} \phi(\mathbf{W}_{q,k} \mathbf{x}_k),\end{aligned}$$

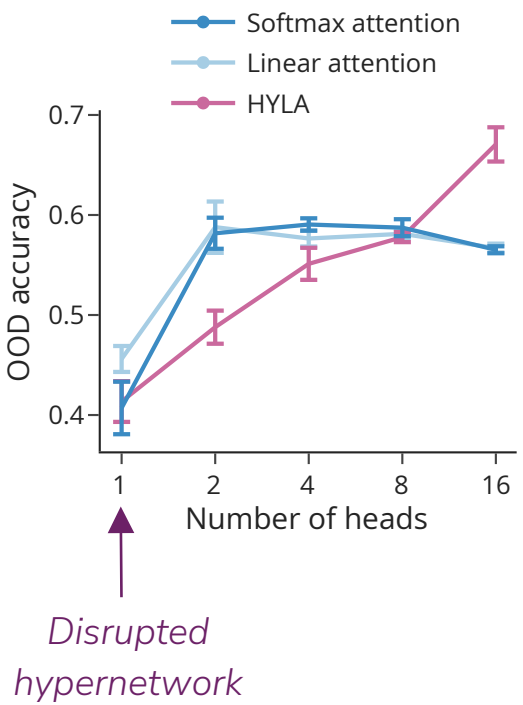
We use HYLA as an experimental intervention to evaluate whether the hypernetwork mechanism is useful for abstract reasoning.

Attention as a hypernetwork

Modifying the hypernetwork mechanism

Reinforcing the hypernetwork mechanism improves compositional generalization, disrupting it hurts compositional generalization.

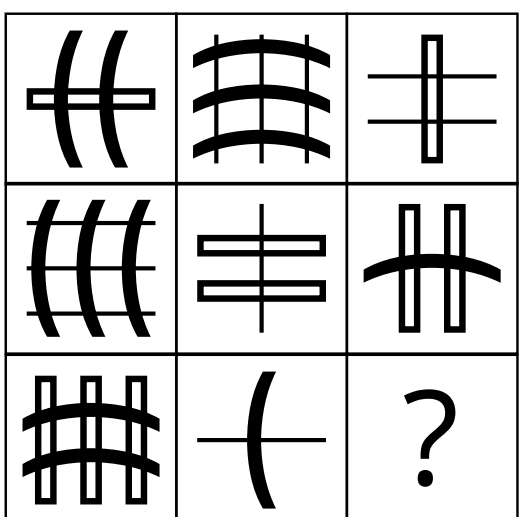




Attention as a hypernetwork

Checkout the full paper if you are curious to see:

- The hypernetwork mechanism at play on fuzzy logic functions
- The impact of the hypernetwork mechanism on language modeling
- The solution to (in Fig. 3B)



Link to OpenReview