# Causality in Neural Networks

## Recurrent independent mechanisms

Maciej Żelaszczyk

December 1, 2021

PhD Student in Computer Science
Division of Artificial Intelligence and Computational Methods
Faculty of Mathematics and Information Science

m.zelaszczyk@mini.pw.edu.pl

Warsaw University of Technology

# Independent mechanisms

Mechanisms:

- Humans are able to adapt to new domains with little to no retraining.
- This might be because we rely on mechanisms that are independent of the particular domain.
- For instance, people are able to recognize distorted images from the get-go.
- It can be hypothesized that these mechanisms are modular, reusable and broadly applicable.

# Independent mechanisms

The *independent mechanisms* (IM) assumption:

- The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

# Independent mechanisms

Let us consider variables $x_1, \ldots, x_d$. If their joint density is Markovian w.r.t. a directed acyclic graph $\mathcal{G}$, we can write:

$$p(\mathbf{x}) = p\left(x_1, \ldots, x_d\right) = \prod_{j=1}^{d} p\left(x_j | \mathrm{pa}_{\mathcal{G}}^{j}\right) \tag{1}$$

where $\mathrm{pa}_{\mathcal{G}}^{j}$ denotes the parents of variable $x_j$ in the graph.

- In the general case, for a given joint density function, we can find many graphs (decompositions) of such form.
- If the edges of $\mathcal{G}$ denote direct causation, then $\mathcal{G}$ is called a *causal* graph and each conditional probability $p\left(x_j | \mathrm{pa}_{\mathcal{G}}^{j}\right)$ can be understood as a *causal mechanism* generating $x_j$ from its parents.
- The presented factorization is a *generative* model in the sense of describing an actual physical *generative* process.

# Independent mechanisms

Consequences of the IM assumption:

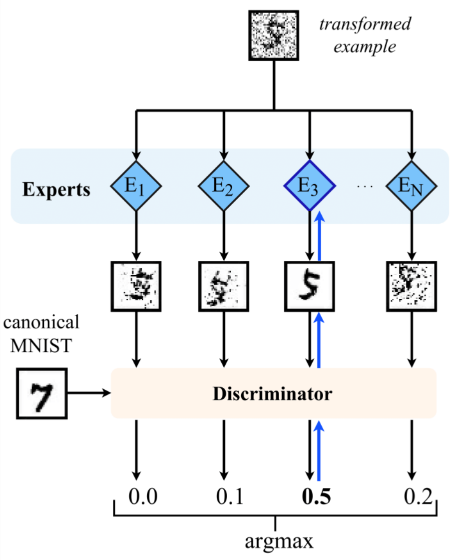
- The causal conditionals are autonomous modules that do not influence or inform each other.
- Knowledge of one mechanism does not contain information about another one.
- Changes in one mechanism do not affect the other mechanisms - *invariance*.
- An intervention on one mechanism does not impact other ones.
- If we change $p\left(x_j|\mathrm{pa}_{\mathcal{G}}^j\right)$, other mechanisms $p\left(x_i|\mathrm{pa}_{\mathcal{G}}^i\right), i \neq j$ do not change.
- Consider that this is not true for other factorizations that do not capture the causal structure.

# Independent mechanisms

Machine learning models expressed in terms of causal mechanisms could:

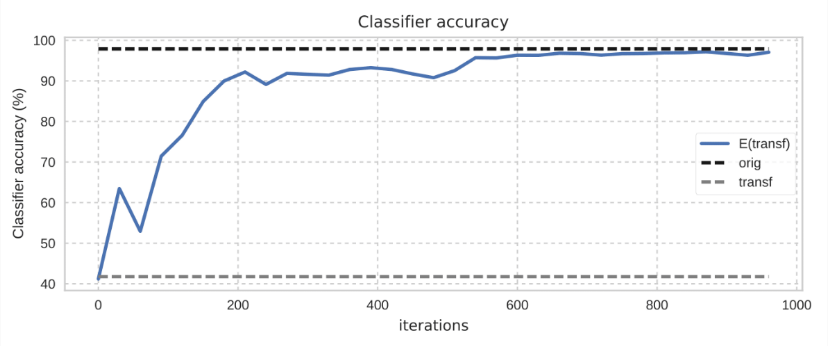
- Facilitate transfer learning, domain adaptation, generalization.
- Provide modularity and the opportunity to train parallel components, which could be recombined into larger systems.
- Offer more interpretability.
- Increase sample efficiency.
- Help in overcoming catastrophic forgetting.

# Inverse mechanisms




Source: [Parascandolo et al., 2018]

# Inverse mechanisms




Source: [Parascandolo et al., 2018]

# Why not one big model?

All that we have covered so far is fine and all but can we just not use one big model to learn the independent mechanisms?

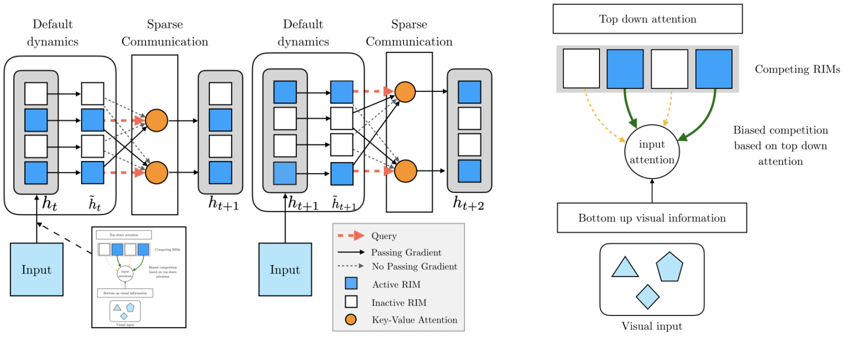
- Consider a simple network.
- Model $k$ independent mechanisms with this net.
- For the hidden states to compartmentalize the different processes, we potentially need to set a portion of weights to 0.
- The fraction to be set to 0 is actually $\frac{k-1}{k}$.
- $\lim_{k \to \infty} \frac{k-1}{k} = 1$.

# Recurrent independent mechanisms

We adopt independent mechanisms to model a recurrent process.

- Divide the model into $k$ modules.
- Each of these modules is recurrent (RIM).
- RIM $k$ at time step has a vector-valued state $h_{t,k}$.
- Parametrized by $\theta_k$ shared across all time steps.
- Individual RIMs compete to process input at time step $t$.
- Only a number of RIMs are activated at each step.
- Sparse communication between RIMs.
- Extensive use of attention.

# Recurrent independent mechanisms




Source: [Goyal et al., 2021]

# Attention

Neural networks are able to operate on sets of typed objects.

- Each *query* represented in a row matrix $Q_{N_r \times d}$.
- $N_r$ - number of queries, $d$ - dimensionality of each query.
- Set of $N_o$ *objects (values)* associated with a *key* matrix $K_{N_o \times d}$, a row matrix of keys.
- Each key is associated with an object (value) $v_i$, which is a row of the value matrix $V_{N_o \times d^*}$.

# Attention

Attention produces combinations of values.

$$\text{attention}(Q, V, K) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

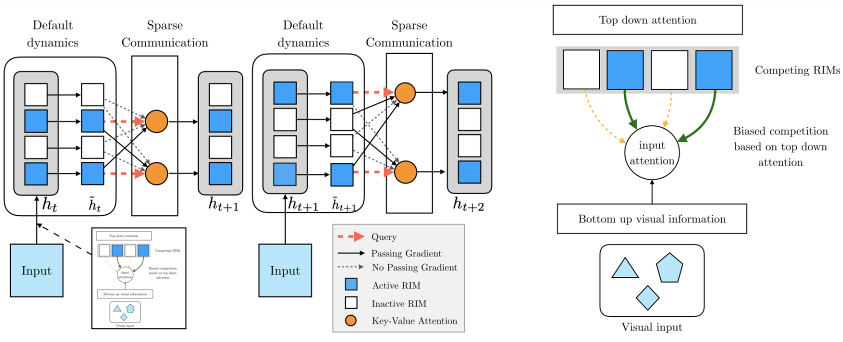
- Softmax applied to rows of $\frac{QK^T}{\sqrt{d}}$.
- Convex combination of the values in the rows of $V$.
- $d$ dimensions can be split into heads with separate attention matrices and write values.

# RIMs as functions

RIMs can operate on values similarly to variables in a programming language:

- Each RIM can be interpreted as a function.
- Values are interchangeable arguments to functions.
- Arguments have a distributed representation for their name (or type) and value.
- Query vector of a RIM specifies the required type.
- RIM applied to a fitting vector.
- Each attention head corresponds to one typed parameter of the function represented by the RIM.
- When the key of an object matches the query of head $k$, it can be used as the $k$-th input vector argument for the RIM.

# Recurrent independent mechanisms




Source: [Goyal et al., 2021]

# Recurrent independent mechanisms

Application of attention in RIMs.

- Multi-head attention for input.
- Input augmented with a zero row.
- Attention calculated for all RIMs.
- Attention scores averaged over heads.
- $k_A$ out of $k$ RIMs with lowest attention scores on the zero row are activated.
- Multi-head attention for communication.
- Attention calculated for active RIMs over all RIMs.

# Copying task


| Copying | $k_{\mathrm{T}}$ | $k_{\mathrm{A}}$ | $h_{\mathrm{size}}$ | **Train(50)** CE | **Test(200)** CE |
|---|---|---|---|---|---|
| **RIMs** | 6 | 4 | 600 | **0.00** | **0.00** |
| | 6 | 3 | 600 | **0.00** | **0.00** |
| | 6 | 2 | 600 | **0.00** | **0.00** |
| | 5 | 2 | 500 | **0.00** | **0.00** |
| **LSTM** | - | - | 300 | 0.00 | 4.32 |
| | - | - | 600 | 0.00 | 3.56 |
| **NTM** | - | - | - | 0.00 | 2.54 |
| **RMC** | - | - | - | 0.00 | 0.13 |
| **Transformers** | | - | - | 0.00 | 0.54 |

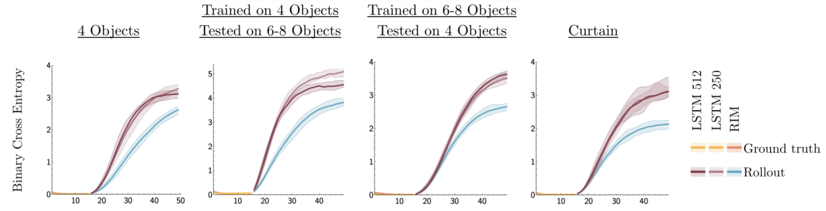
Source: [Goyal et al., 2021]

# Sequential MNIST

| Sequential MNIST | $k_T$ | $k_A$ | $h_{size}$ | 16 x 16 Accuracy | 19 x 19 Accuracy | 24 x 24 Accuracy |
|---|---|---|---|---|---|---|
| RIMs | 6 | 6 | 600 | 85.5 | 56.2 | 30.9 |
| | 6 | 5 | 600 | 88.3 | 43.1 | 22.1 |
| | 6 | 4 | 600 | **90.0** | **73.4** | **38.1** |
| LSTM | - | - | 300 | 86.8 | 42.3 | 25.2 |
| | - | - | 600 | 84.5 | 52.2 | 21.9 |
| EntNet | - | - | - | 89.2 | 52.4 | 23.5 |
| RMC | - | - | - | 89.58 | 54.23 | 27.75 |
| DNC | - | - | - | 87.2 | 44.1 | 19.8 |
| Transformers | - | | - | **91.2** | 51.6 | 22.9 |

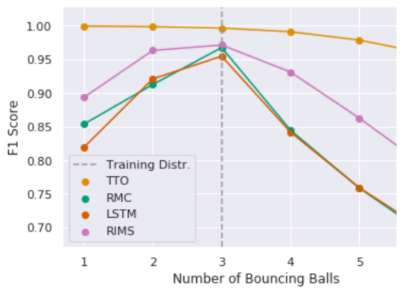
Source: [Goyal et al., 2021]

# Bouncing balls environment



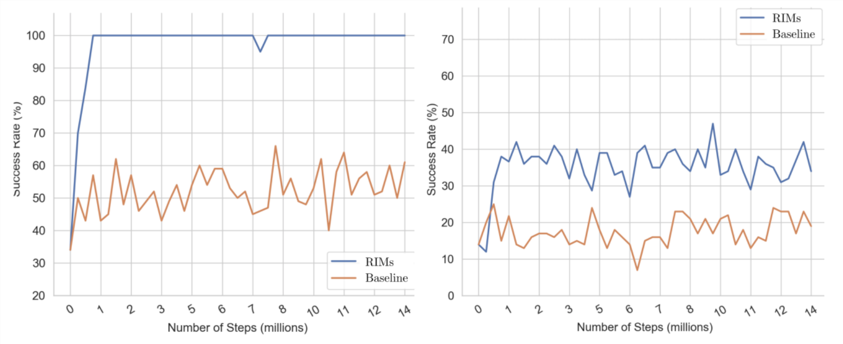

Source: [Goyal et al., 2021]

# Bouncing balls environment



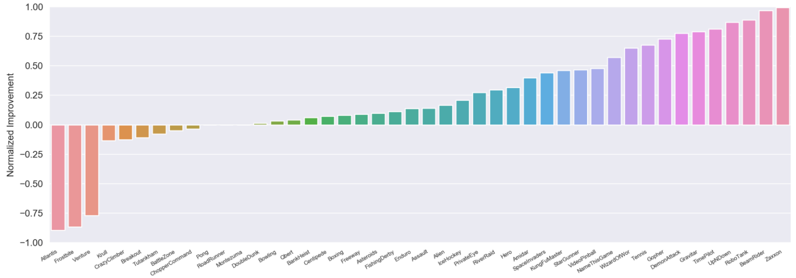

Source: [Goyal et al., 2021]

# Robustness to distractors




Source: [Goyal et al., 2021]

# Atari




Source: [Goyal et al., 2021]

## General results

Ablations:

- Sparse activation is necessary, but works for a wide range of hyperparameters.
- Input-attention is necessary.
- Communication between RIMs improves performance.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. (2021).
**Recurrent independent mechanisms.**
In *9th International Conference on Learning Representations (ICLR)*.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., and Schölkopf, B. (2018).
**Learning independent causal mechanisms.**
In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4036–4044, Stockholmsmässan, Stockholm Sweden. PMLR.