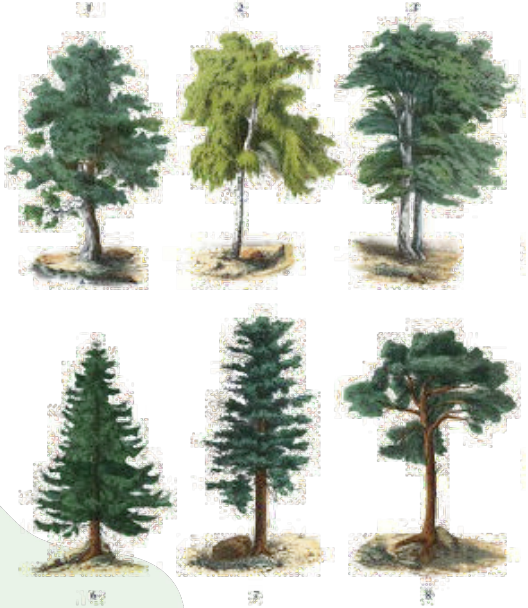
Attentional Meta-learners for Few-shot Polythetic Classification

me

Ben Day*, Ramon Viñas*, Nikola Simidjievski, Pietro Liò

equal contribution

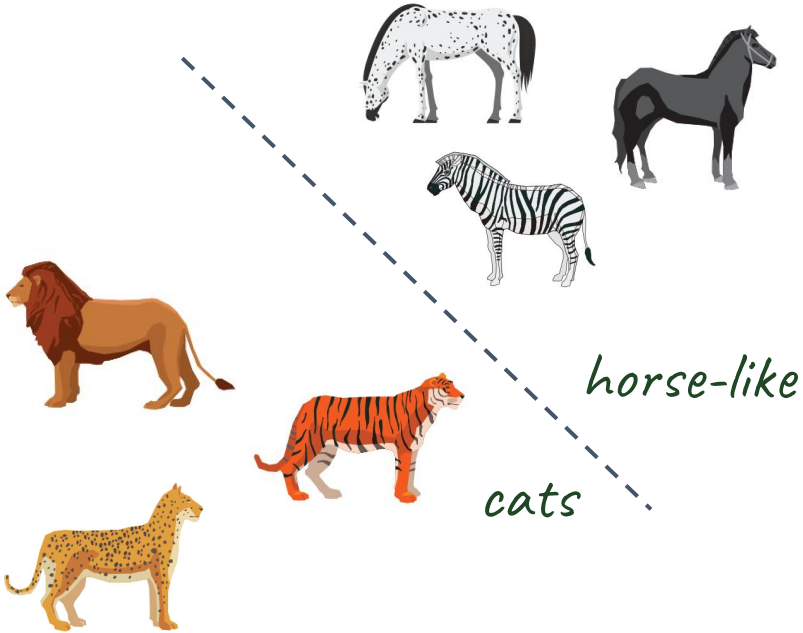
Typically: few-shot means unseen classes

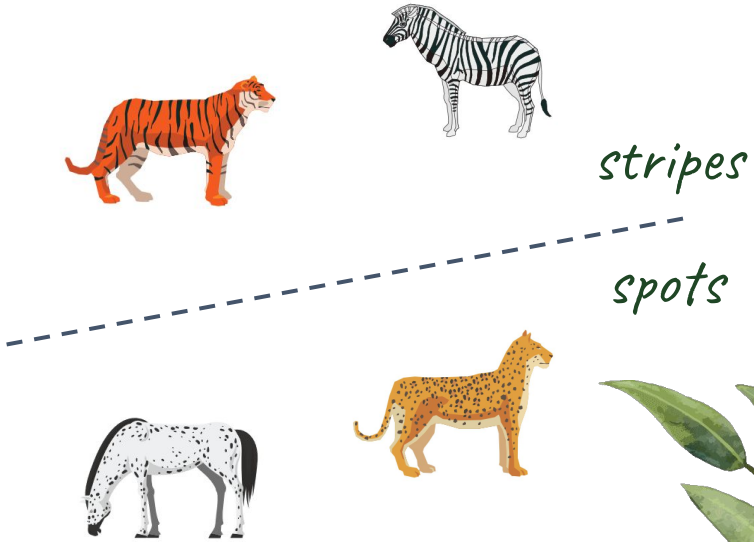


train

test

Alternative: unseen ways of categorising





This work

> Relate the new framing to existing ideas

> Apply standard approaches

> Quantify limitations through well-defined abstract problems

> Propose a simple mechanism based on self-attention

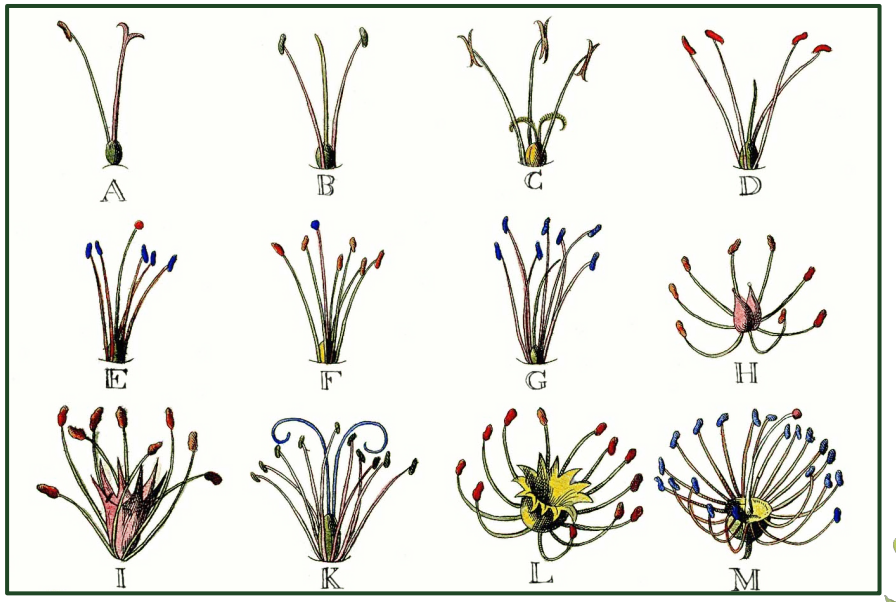
> Present experiments on large-scale real-world problems

Existing ideas: Monothetic / Polythetic

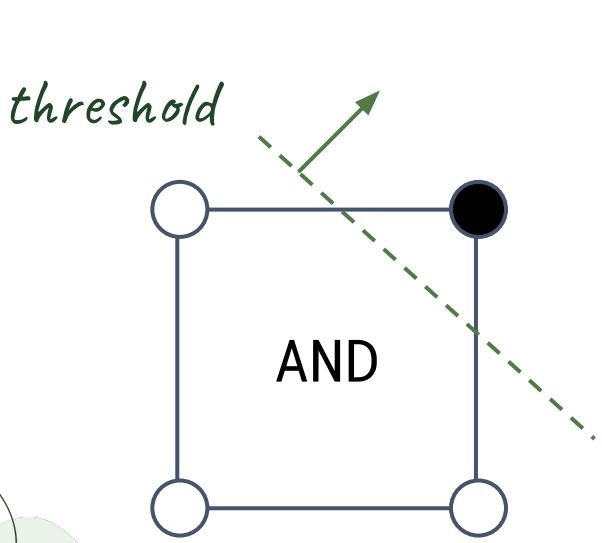
(at least) one feature is sufficient and necessary

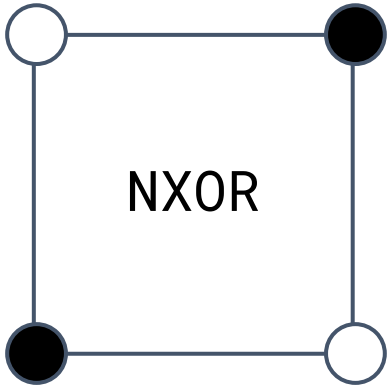
vs

combinations where none are sufficient or necessary



Existing ideas: Boolean functions

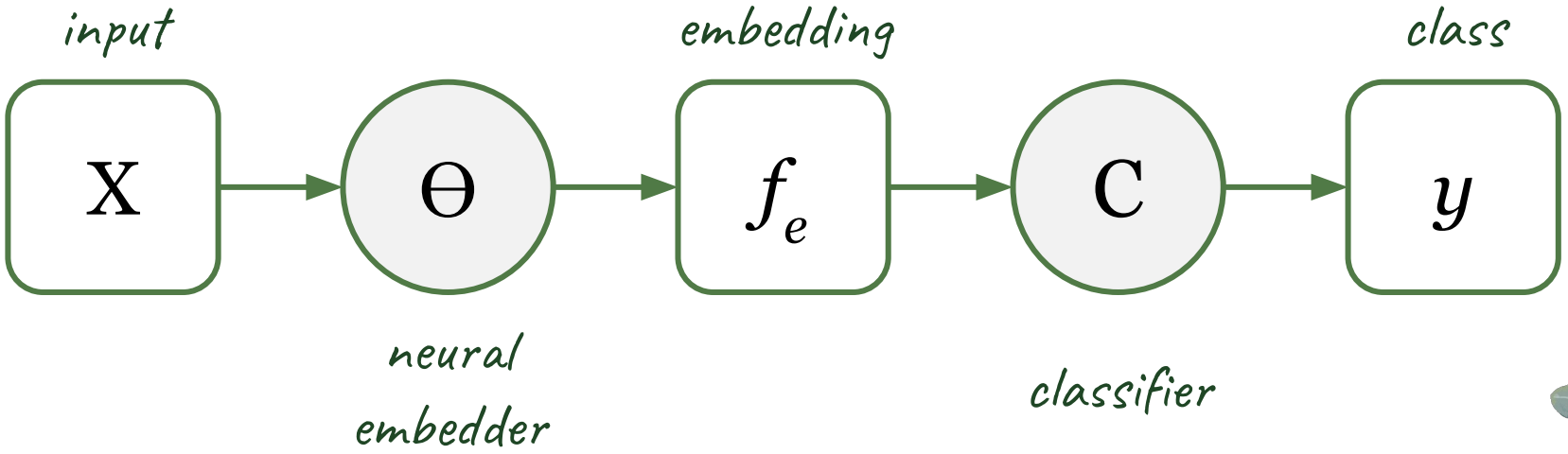




non-threshold

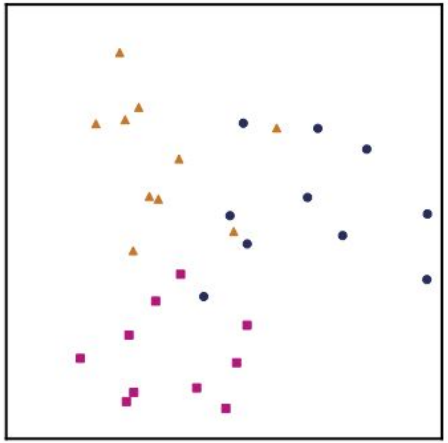
$$2^{n^2} \ll 2^{2^n}$$

Metric-based meta-learners

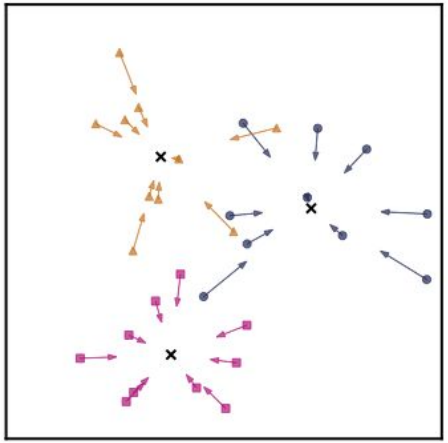


Metric-based meta-learners

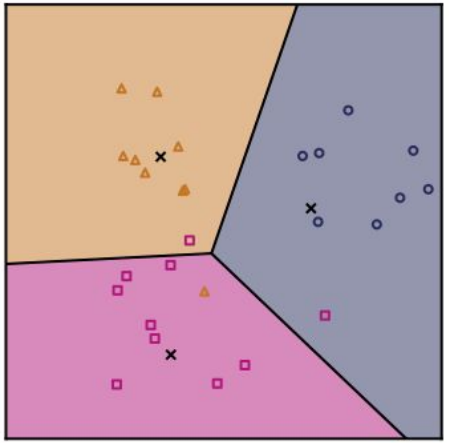
Prototypical Networks



labelled examples



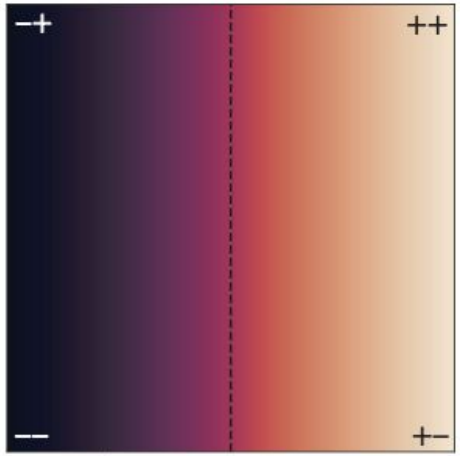
prototypes

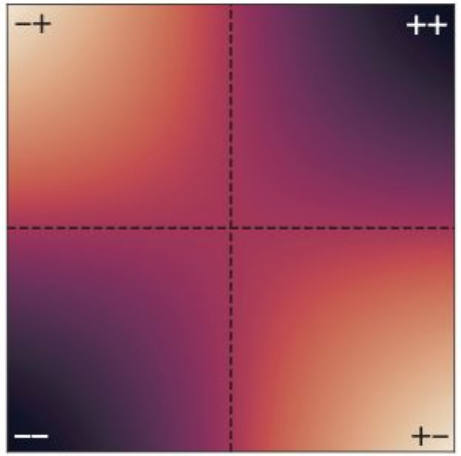


boundaries

Metric-based meta-learners

Matching Networks

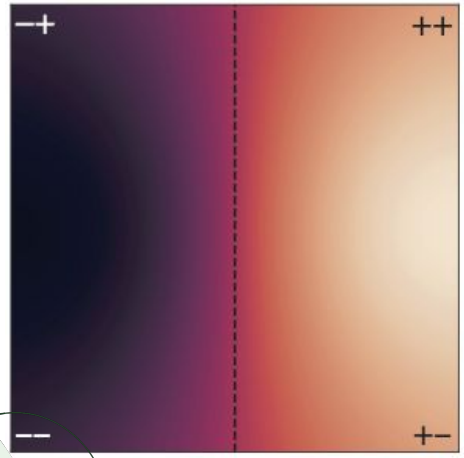




$$\hat{y} = \sum a(\hat{x}, x_i) y_i$$

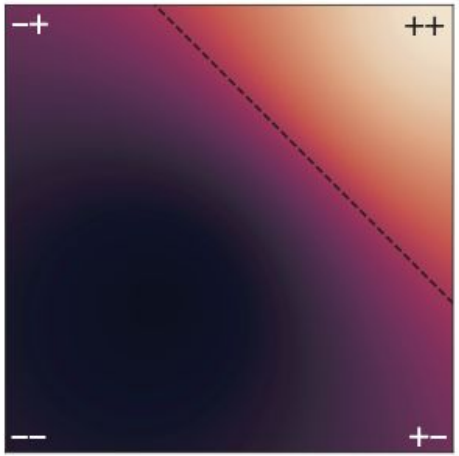
Meta-learning Boolean functions

$f_1(x, y) = x$



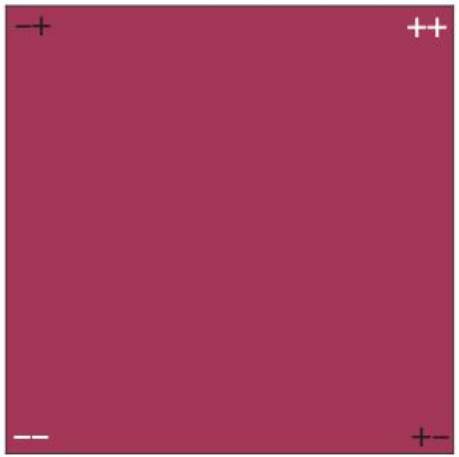
Prototype, $x = 0$

$f_2(x, y) = \text{AND}(x,y)$



Prototype, $y = -x + 2/3$

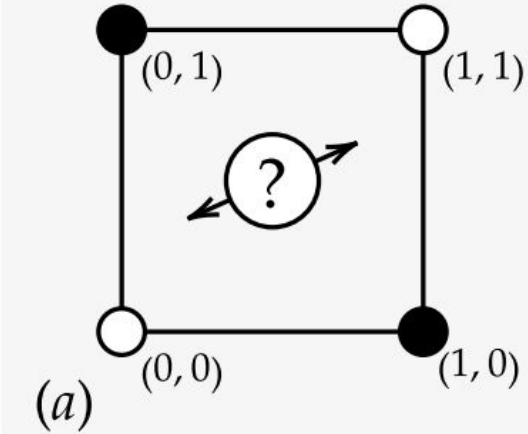
$f_3(x, y) = \text{XOR}(x,y)$

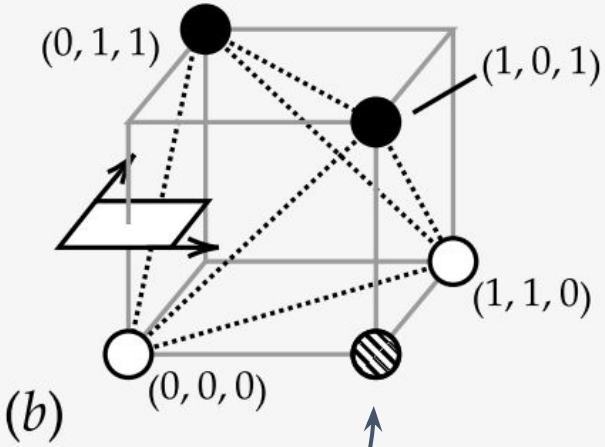


Prototype, undefined

not good

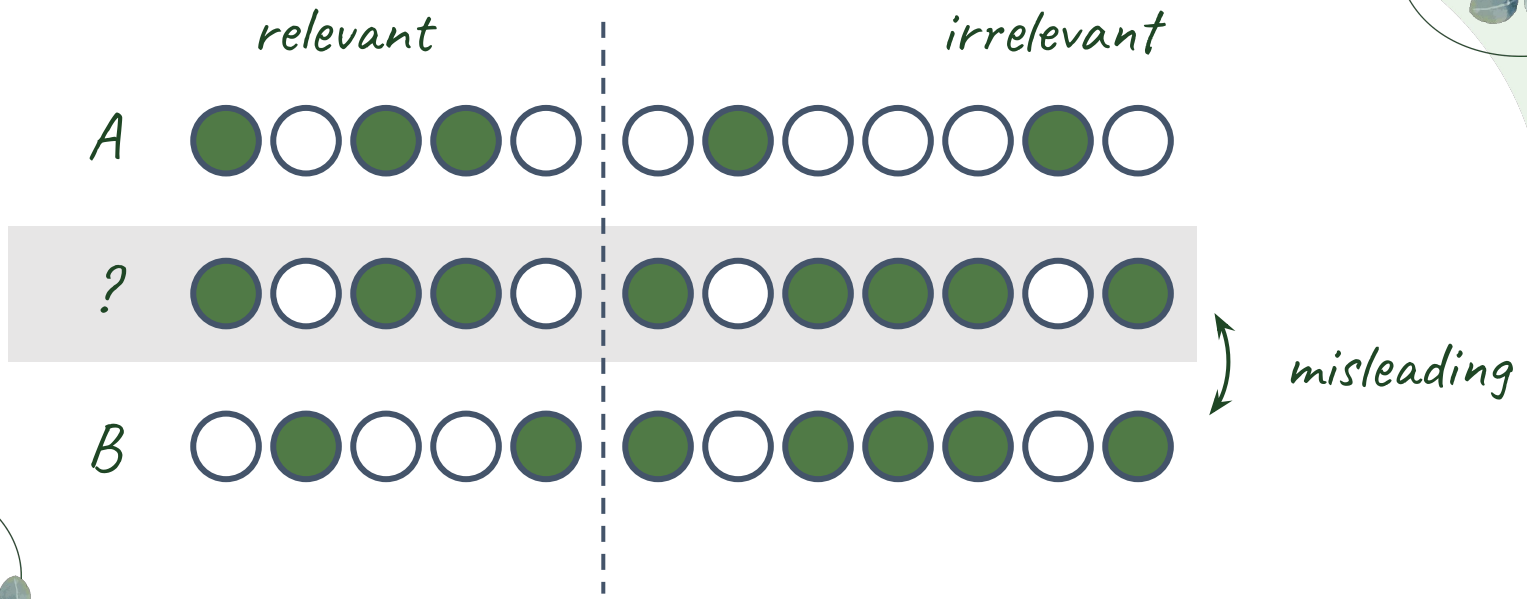
Meta-learning Boolean functions



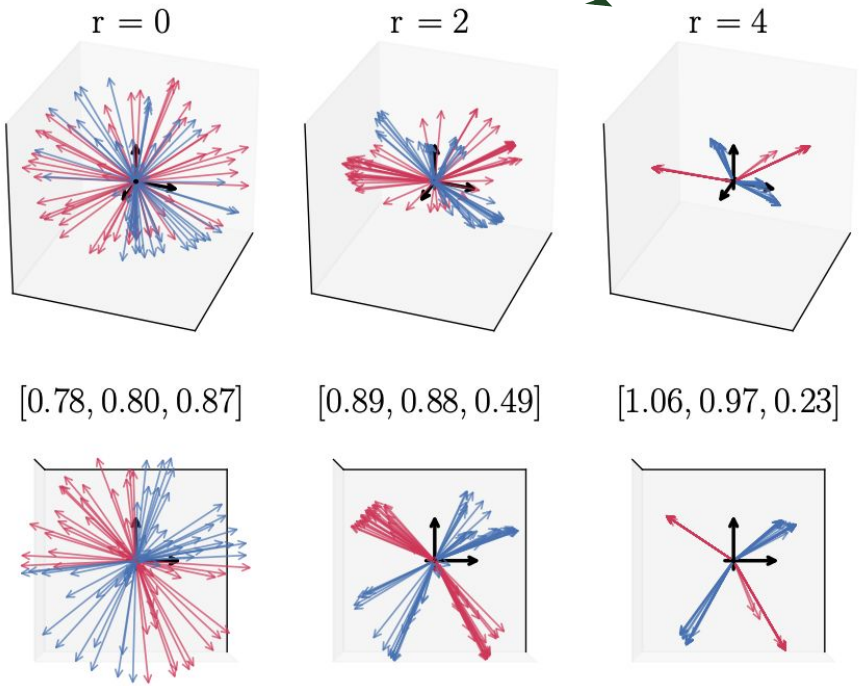


pseudo-feature

Meta-learning Boolean functions



irrelevant z-dimension decays



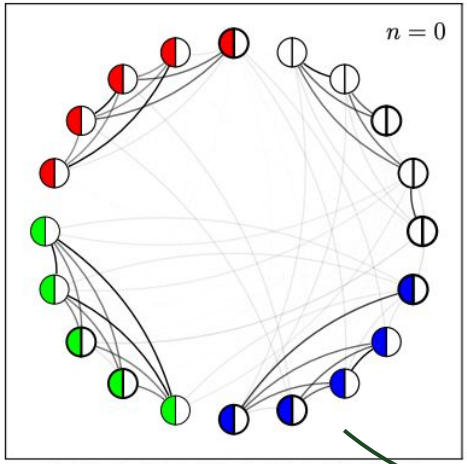
Our proposal

3: $\mathbf{x}_i \leftarrow (\mathbf{x}_i - \mu_X)/(\sigma_X + \epsilon)$ {standardise}
4: **for** $r \leftarrow 1$ **to** R **do**
5: **for** $k \leftarrow 1$ **to** K **do**
6: $\mathbf{X}_k \leftarrow \text{softmax}(\tau \mathbf{X}_k \mathbf{X}_k^T)\mathbf{X}_k$ {softmax is row-wise}
7: **end for**
8: **end for**
9: $\boldsymbol{f} \leftarrow \text{dispersion}(\{\mathbf{x}_i\}_{i \in S_i})$ {mean-absolute-deviation, std. dev etc.}

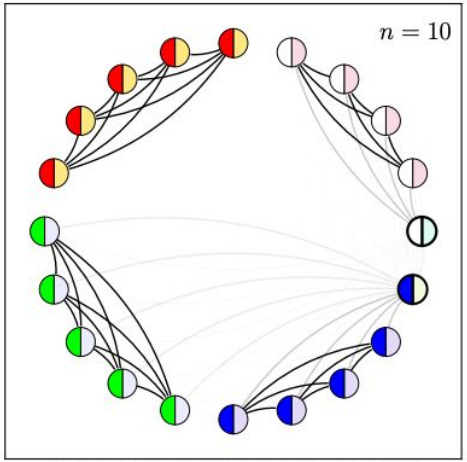
self-attend within classes

irrelevant features decay

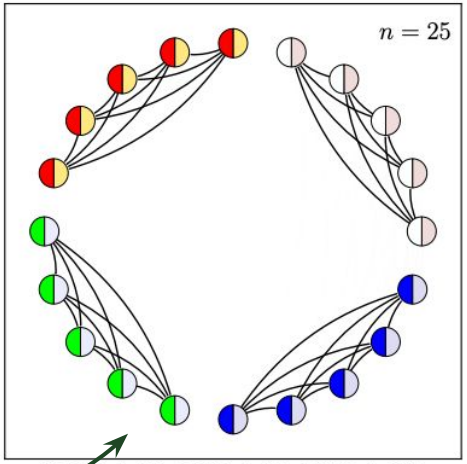
Our proposal



[1.00, 1.00, 1.00, 1.00, 1.00, 1.00]



[1.00, 1.00, 1.00, 0.89, 0.80, 0.69]

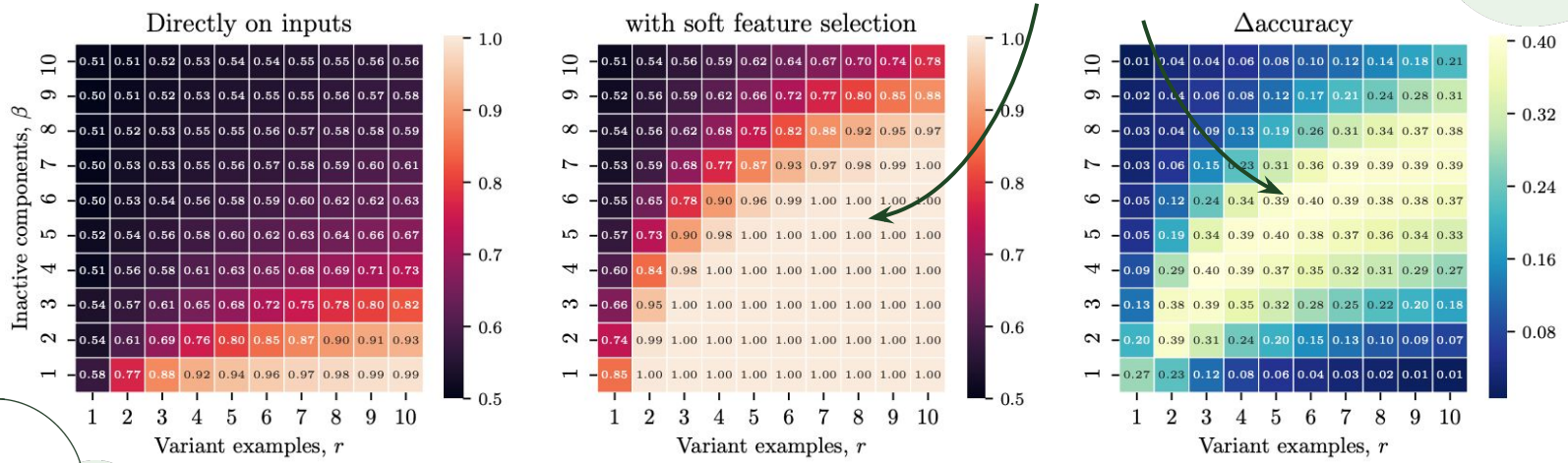


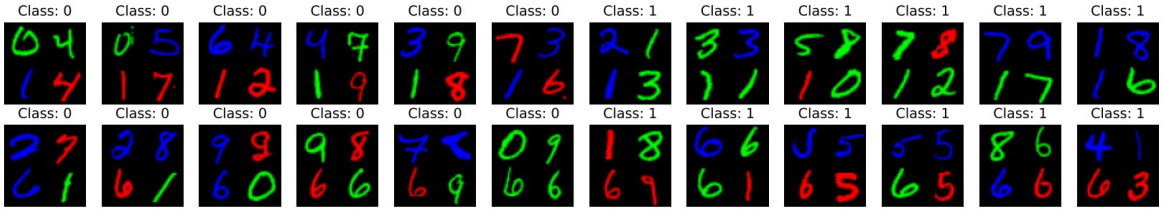
[1.00, 1.00, 1.00, 0.87, 0.78, 0.66]

self-organisation

Synthetic experiments

boost to performance





Real-world experiments

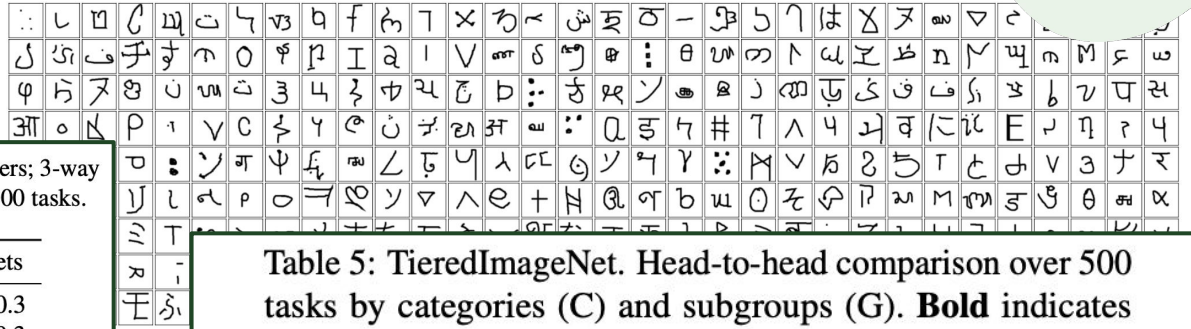


Table 3: Omniglot. 20-way, 5-shot characters; 3-way alphabets. Mean and standard error on 1000 tasks.

	Characters	Alphabets
PN	**98.6 ± 0.0**	83.4 ± 0.3
MN	91.1 ± 0.1	78.4 ± 0.3
FS+MN	96.2 ± 0.0	94.2 ± 0.2
IMP	**98.6 ± 0.0**	**96.0 ± 0.2**
MAML	94.0 ± 0.1	89.9 ± 0.3
MN*	97.9 ± 0.1	81.3 ± 0.3
NN*	98.3 ± 0.0	95.7 ± 0.3
FS+MN*	98.1 ± 0.0	**96.0 ± 0.2**
FS+NN*	98.3 ± 0.0	**96.0 ± 0.2**

Table 5: TieredImageNet. Head-to-head comparison over 500 tasks by categories (C) and subgroups (G). **Bold** indicates significance at the $p < 0.001$ level.

			C = 2		C = 4		C = 8	
G	X	Y	X / Y	(tie)	X / Y	(tie)	X / Y	(tie)
5	FS	MN	**230** / 162	(108)	**329** / 101	(70)	**343** / 113	(44)
	FS	PN	**319** / 136	(45)	**339** / 142	(19)	239 / 247	(14)
10	FS	MN	**258** / 185	(57)	**357** / 108	(35)	**413** / 68	(19)
	FS	PN	**374** / 101	(25)	**396** / 99	(5)	**296** / 191	(13)

So much to talk about, so little time

> Lots of connections to related work

> Greater depth on the connections and limitations

> Additional experiments

> The code is on github

Thanks !

Attentional Meta-learners for Few-shot Polythetic Classification

`github.com/rvinas/polythetic_metalearning`

Ramon **rv340@cam.ac.uk**

Ben **bjd39@cam.ac.uk**

Nikola **ns779@cam.ac.uk**

Pietro **pl219@cam.ac.uk**