Structure Assembly in Knowledge Base Representation

Matthias Lalisse

The “Language of Thought” Hypothesis

Classical cognitive science:

> *Cognitive capacities are systems of computational procedures that operate over domains of symbols to produce behavior*

i.e. cognition in general has the formal structure of *language*

The Fodor & Pylyshyn Formula

Higher-order cognition is:

Productive: In certain (but not all) domains, there is “discrete infinity”

Systematic: Cognitive representations are systematically linked to one another in virtue of what constituents appear in them

Roughly, algebraic closure of the alphabet under the operations of the “grammar”: if *Mary loves Kevin* is a sentence, then *Kevin loves Mary* is also a sentence

Compositional: There are *semantic* relations between representations that depend on the constituents appearing in them

E.g.

$$(\texttt{cat},\ \texttt{has\_part},\ \texttt{paw})$$
$$(\texttt{panther},\ \texttt{is\_instance},\ \texttt{cat})$$
$$\implies (\texttt{panter},\ \texttt{has\_part},\ \texttt{paw})$$

Implications of F&P

- Cognitive theories ought to be able to satisfy F&P's "benchmarks"
- They go further & conjecture that any cognitive theory that satisfies the "benchmarks" are necessarily isomorphic to those systems

Questions raised for neural models:

- How would these symbolic systems be realized in neural models? (the **Implementationalist Question**)
- Are there phenomena that symbolic theories do not cover, or that are more cumbersome for them to cover relative to non-symbolic alternatives? (the **Symbolic Describability Question**)
  - E.g. similarity relations, analogies, prototype effects, etc

Roadmap

- Connectionist solutions to the Fodor & Pylyshyn criteria
- Properties of some binding operators
- Quasi-compositional phenomena
- Harmony Maximization: a framework for noncompositional computation
- 3 models:
  - Gradient Graphs
  - Harmonic Memory Networks
  - Spatial Attention Networks

Symbolic systems in neural systems

Classical responses to the F&P framework: Provide explicit mechanisms that satisfy the three criteria

The goal: provide explicit mechanisms that account for the F&P properties

Vector Symbolic Architectures

Proposals for systems that operate over vectors and derive the F&P properties

General framework:
There are sets of symbols (fillers) and roles, and a binding operation that combines them into pairwise associations

$$\begin{aligned} \text{Binding operator:} \quad & \mathbb{B}(\boldsymbol{x}, \boldsymbol{y}) \\ \text{Unbinding operator:} \quad & \mathbb{U}(\boldsymbol{x}, \mathbb{B}(\boldsymbol{x}, \boldsymbol{y})) \approx \boldsymbol{y} \end{aligned}$$

There is a coupled unbinding operator that is used to extract parts of the assembled structure

Add appropriate algorithms and:
⇒ Yields the Language of Thought properties

Binding models

- **Tensor Product Representations/TPRs** (Smolensky 1990, applied in e.g. Schlag 2018)
  - Binding: tensor product
  - Unbinding: dot product with structural role vectors
  - Gives **exact retrieval** of the vector associations but in a large representation

$$\boldsymbol{r} \cdot (\boldsymbol{r} \otimes \boldsymbol{x}) = \boldsymbol{x}$$

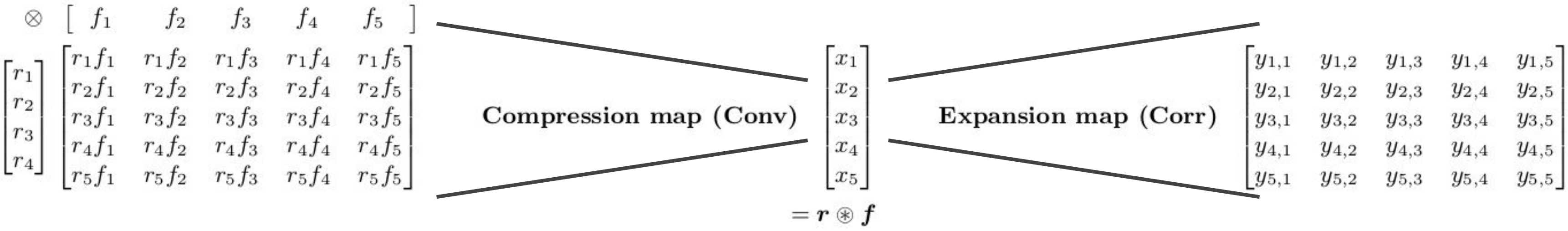
- **Holographic Reduced Representations/HRRs** (Plate 1995, applied in e.g. Nickel 2015, NENGO)
  - Binding: circular convolution
  - Unbinding: circular correlation
  - A kind of “compressed” tensor product
  - The binding has the same dimension as the inputs, but recovery is only approximate

$$\boldsymbol{r} \star (\boldsymbol{r} \circledast \boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\varepsilon} \approx \boldsymbol{x}$$

Noise term (HRR approx)

On the relation between HRRs and TPRs

- Why are HRRs "good" binding mechanisms?
  - **Theorem**: The circular correlation tensor is the Moore-Penrose inverse of the circular convolution tensor.
    - **Corollary**: Correlation provides an *optimal reconstruction* of a TPR that is encoded into a smaller space by the convolution tensor



w.r.t. Convolution, Correlation minimizes the expected retrieval error:

$$\mathbb{E}\left[|| \boldsymbol{r} \otimes \boldsymbol{f} - \mathrm{Corr}\left(\mathrm{Conv}\left(\boldsymbol{r} \otimes \boldsymbol{f}\right)\right) || \right]$$

- **HRR computation stream**:
  - Take the TPR of a structure that is bound
  - Compress the TPR using the forward map (convolution)
  - Retrieve the *optimal approximation* of the original TPR using the correlation map
  - Do standard standard TPR operations (unbinbing using dot product) to process the structure

Quasi-compositional phenomena

- Copredication:
  - Dinner was tasty but took forever.
    - [**Dinner**$_{\text{substance}}$] was tasty but [**dinner**$_{\text{event}}$] took forever
- Coercion:
  - Julie enjoyed the book.
    - ⇒ Julie enjoyed reading **the book**.
  - The goat enjoyed the book.
    - ⇒ The goat enjoyed eating **the book**.

Physical substance type

Informational content type

Event type

adapted from (Asher 2011)

Harmony Maximization: “supracompositional” computational component

Cognitive representations resemble a “Language of Thought” as a first approximation

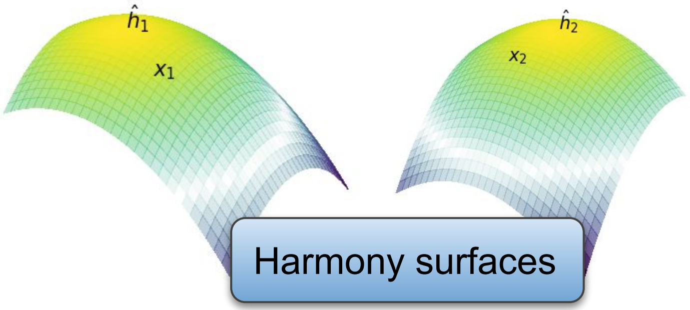
- Core compositional operations take constituents of a structure and combine them using systematic operations
- A recurrent neural network optimizes the representation on the basis of a Harmony function

Compositional input rep

$$\mathcal{H}(\boldsymbol{h}, \boldsymbol{x}) = \frac{1}{2}\left[\boldsymbol{h}^\top \mathbb{W} \boldsymbol{h} + \boldsymbol{b}^\top \boldsymbol{h} - \lambda(\boldsymbol{h} - \boldsymbol{x})^\top (\boldsymbol{h} - \boldsymbol{x})\right]$$

Core Harmony Faithfulness

Optimize this (hidden layer state vector)

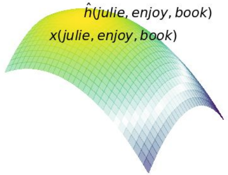


“Books” in an HMax network

$$(\texttt{julie}, \texttt{enjoy}, \texttt{book}) \xrightarrow{\text{embed}} f_{\text{comp}}(\boldsymbol{e}_{\texttt{julie}}, \boldsymbol{r}_{\texttt{enjoy}}, \boldsymbol{e}_{\texttt{book}})$$

$$\xrightarrow{\text{compose}} \boldsymbol{x}_{(\texttt{julie},\texttt{enjoy},\texttt{book})}$$

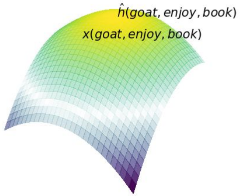
$$\xrightarrow{\text{HMax}} \hat{\boldsymbol{h}}_{(\texttt{julie},\texttt{enjoy},\texttt{book})} \approx \text{"Julie read the book and liked it"}$$

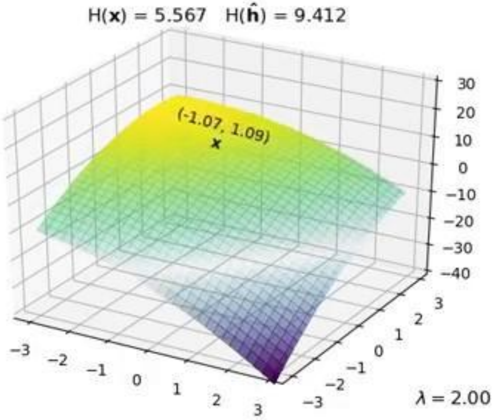


$$(\texttt{goat}, \texttt{enjoy}, \texttt{book}) \xrightarrow{\text{embed}} f_{\text{comp}}(\boldsymbol{e}_{\texttt{goat}}, \boldsymbol{r}_{\texttt{enjoy}}, \boldsymbol{e}_{\texttt{book}})$$

$$\xrightarrow{\text{compose}} \boldsymbol{x}_{(\texttt{goat},\texttt{enjoy},\texttt{book})}$$

$$\xrightarrow{\text{HMax}} \hat{\boldsymbol{h}}_{(\texttt{goat},\texttt{enjoy},\texttt{book})} \approx \text{"The goat ate the book and liked it"}$$

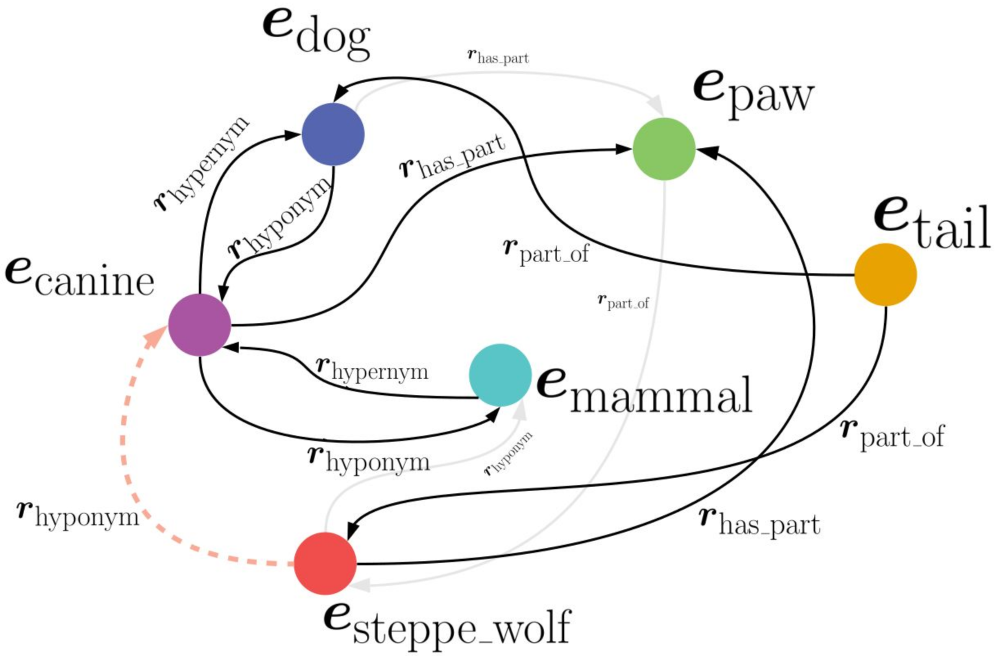




Problem Domain:

Knowledge Base Completion

Take a database of facts and generalize the database to new facts



(dog, has_part, paw)
(dog, hyponym, canine)
(tail, part_of, dog)
(canine, hypernym, dog)
(mammal, hypernym, canine)
(steppe_wolf, has_part, paw)
~~(canine, hypernym, steppe_wolf)~~
(tail, part_of, steppe_wolf)
(tail, part_of, dog)

Infer:

⟹ (steppe_wolf, hyponym, canine)

Generic strategy: Embed entities and relations, and design a function that takes the embeddings & combines them systematically to derive a score

⇒ Removing this premise makes the inference nondeductive

(Lalisse & Smolensky 2019)

Gradient Graphs

Application of the mechanisms of Harmonic Grammar (compositional assembly + optimization of the compositional representation) to KBC

Basic proposal:

Use an array of **composition functions** to build representations of knowledge base entries

Augment the compositional representations with a **semantic optimization function** that subjects the compositional representations to learned constraints

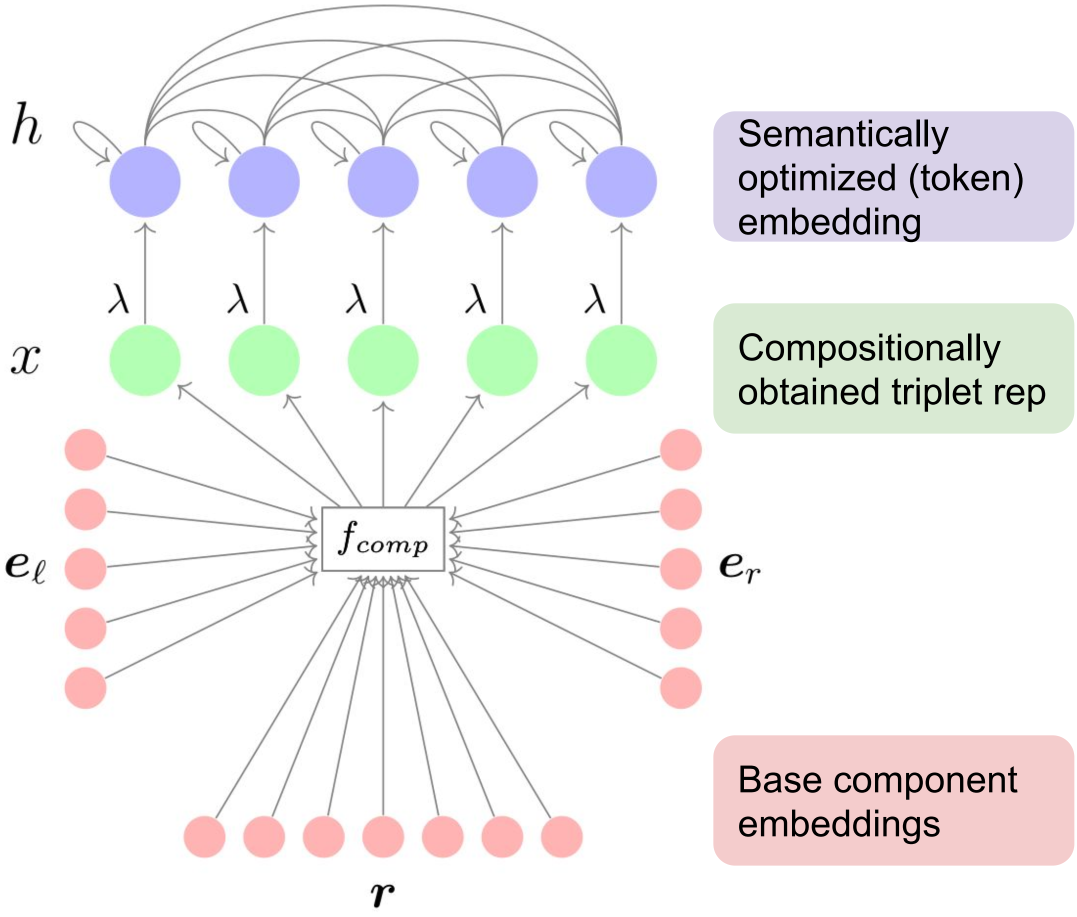
Gradient Graph Network

Three-layer neural model:

Embedding layer

Feedforward **composition layer**

Recurrent **optimization layer**



GG Composition Functions

Three multilinear functions of the entity & relation embeddings

Harmonic Tensor Product Representations

$$\boldsymbol{x}_{\text{HTPR}} = \boldsymbol{e}_\ell \otimes \boldsymbol{r} \otimes \boldsymbol{e}_r$$

$$[\boldsymbol{x}]_{ijk} = [\boldsymbol{e}_\ell]_i [\boldsymbol{r}]_j [\boldsymbol{e}_r]_k$$

Harmonic Elementwise Multiplication (DistMult in Wang 2015)

$$\boldsymbol{x}_{\text{HDM}} = \boldsymbol{e}_\ell \odot \boldsymbol{r} \odot \boldsymbol{e}_r$$

$\odot$: elementwise multiplication

Harmonic Circular Correlation (HolE in Nickel 2015)

$$\boldsymbol{x}_{\text{HHoLE}} = \boldsymbol{r} \odot (\boldsymbol{e}_\ell \star \boldsymbol{e}_r)$$

$$[\boldsymbol{e}_\ell \star \boldsymbol{e}_r]_j = \sum_i [\boldsymbol{e}_\ell]_i [\boldsymbol{e}_r]_{(i+k) \bmod d}$$

(circular correlation)

Results

Tensor Product Representations

$$x_{\text{HTPR}} = \boldsymbol{e}_\ell \otimes \boldsymbol{r} \otimes \boldsymbol{e}_r$$

Un-optimized (purely compositional)

λ	MR	MRR	H@1	H@3	H@10
∞	150	.278	.192	.305	.447
1.0	**134**	**.295**	**.204**	**.326**	**.471**

TPRs: Opt > No-opt

DistMult: Elementwise multiplication (Yang 2015/ Kaldec 2017)

$$x_{\text{HDM}} = \boldsymbol{e}_\ell \odot \boldsymbol{r} \odot \boldsymbol{e}_r$$

HHolE/Correlation (Nickel 2016)

$$x_{\text{HHoLE}} = \boldsymbol{r} \odot (\boldsymbol{e}_\ell \star \boldsymbol{e}_r)$$

	FB15K						WN18					
		Rank		Hits@				Rank		Hits@		
Model	λ	MR	MRR	1	3	10	λ	MR	MRR	1	3	10
DistMult	-	-	.350	-	-	.577	-	-	.830	-	-	.942
Ensemble DM†	-	36	**.837**	**.797**	-	**.904**	-	457	.790	**.784**	-	.950
DistMult*	-	28	.710	.605	.792	.876	-	220	.825	.714	.938	.950
HDistMult	∞	**23**	.806	.751	**.845**	.898	∞	**164**	**.841**	.740	**.943**	**.955**
HDistMult	50.0	**23**	.742	.661	.799	.881	3.0	184	.831	.732	.931	.945
HolE	-	-	.524	.402	.613	.739	-	-	.938	.930	**.945**	.949
HolE*	-	39	.409	.289	.464	.647	-	205	.916	.893	.936	.946
HHolE	∞	32	.682	.575	.763	.850	∞	293	.919	.903	.934	.942
HHolE	1.0	**21**	**.796**	**.727**	**.848**	**.901**	2.0	**183**	**.939**	**.931**	**.945**	**.951**

HRRs: Opt > No-opt

HRRs: Best models overall

Optimized triplets

Obama starts close to 2 Dem Senators, his SecState

Post-opt: only presidents

US Presidents

	George W. Bush			Barack Obama	
n	x (compositional)	$\hat{h}$ (optimized)	n	x (compositional)	$\hat{h}$ (optimized)
1	**George H. W. Bush**	**George H. W. Bush**	1	Hillary Rodham Clinton	**George W. Bush**
2	**Bill Clinton**	**Bill Clinton**	2	Al Gore	**Bill Clinton**
3	**Jimmy Carter**	**Jimmy Carter**	3	**George W. Bush**	**John F. Kennedy**
4	**John F. Kennedy**	**Ronald Reagan**	4	**Bill Clinton**	**Ronald Reagan**
5	**Ronald Reagan**	**Barack Obama**	5	**John F. Kennedy**	**George H. W. Bush**

	John McCain			Al Gore	
n	x (compositional)	$\hat{h}$ (optimized)	n	x (compositional)	$\hat{h}$ (optimized)
1	John Kerry	John Kerry	1	**Barack Obama**	Condoleezza Rice
2	Hillary Rodham Clinton	Colin Powell	2	**George W. Bush**	John C. Calhoun
3	Colin Powell	Nancy Pelosi	3	Colin Powell	Colin Powell
4	**Richard Nixon**	Joe Biden	4	Condoleezza Rice	Hillary Rodham Clinton
5	**Herbert Hoover**	Dick Cheney	5	**John F. Kennedy**	John Kerry

Gore starts close to presidents

No presidents

Optimized triplets

Already prototypical example

Neighborhood stays the same

Guises of Bob Dylan

	Singer-Songwriter			Screenwriter	
n	x (compositional)	$\hat{h}$ (optimized)	n	x (compositional)	$\hat{h}$ (optimized)
1	**Eric Clapton**	**Bonnie Raitt**	1	**John Lennon**	**John Lennon**
2	**Bonnie Raitt**	**Eric Clapton**	2	Jimi Hendrix	**Barbara Streisand**
3	**Van Morrison**	**Van Morrison**	3	**Barbara Streisand**	**Eric Idle**
4	**B.B. King**	**B.B. King**	4	Eric Clapton	**Nick Cave**
5	**Bob Seger**	**Bob Seger**	5	Eddie Vedder	**Alan Bergman**

	Disc Jockey			Writer	
n	x (compositional)	$\hat{h}$ (optimized)	n	x (compositional)	$\hat{h}$ (optimized)
1	Tom Petty	**Steven Van Zandt**	1	**John Lennon**	**Alanis Morissette**
2	Warren Zevon	**Erykah Badu**	2	**Alanis Morissette**	**John Lennon**
3	Willie Nelson	**Alice Cooper**	3	Paul McCartney	**Leonard Cohen**
4	John Mayer	John Mayer	4	Tina Turner	**Leonard Bernstein**
5	Steve Earle	**Moby**	5	Dolly Parton	**Prince**

Moves nearer to Musicians-who-were-also-DJs

Harmonic Memory Networks

With Paul Smolensky
& Eric Rosen

In GGs, we took the representations of constituents to be atomic (i.e. there is no explicit internal structure to the learned embeddings)

Harmonic Memory Networks introduce compositional structure directly into the embeddings

The framework: Entities are represented as **memory states**

Harmonic Memory Networks

Gradient Graphs: Compositionality + HMax, but representations of constituents are treated as atomic

Harmonic Memory Networks: Add compositional structure to the representations of the entities themselves using filler-role binding operations

> Framework: Entities are represented as **memory states** composed of pairwise bindings of entity and relation vectors.
>
> Related to Graph Convolution methods (Shichtkrull 2017, Dettmers 2018) and recent Graph Attention Networks (Nathani 2019)

Representing Entities

Target: a memory state that includes all the links relevant to a given query

Scoring function for each neighborhood link, with the function depending on the query

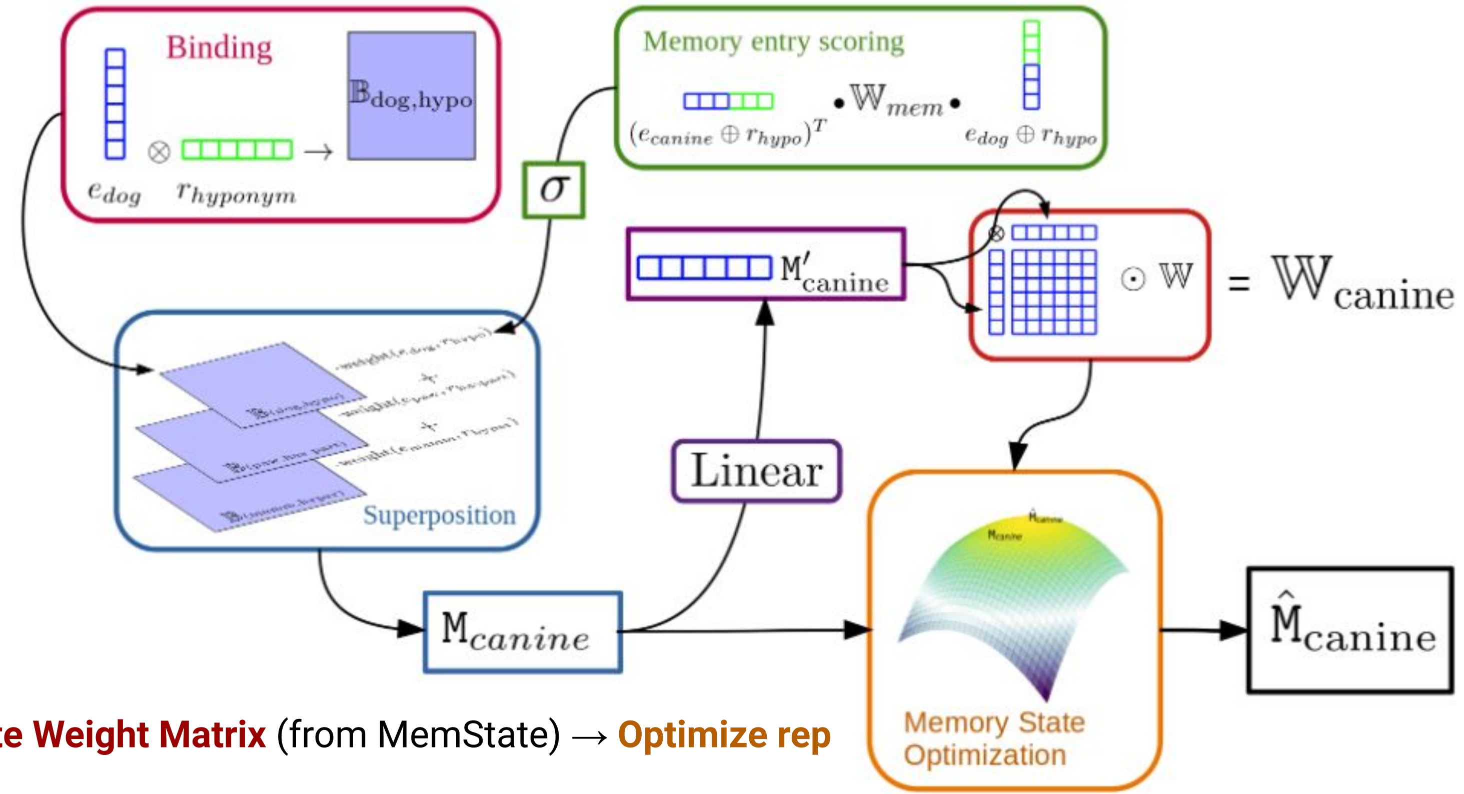
$$\text{weight}(\boldsymbol{e}_c, \boldsymbol{r}_c | \boldsymbol{e}_i, \boldsymbol{r}_q) = \sigma\left((\boldsymbol{e}_i \oplus \boldsymbol{r}_q)^\top W_{\text{score}} (\boldsymbol{e}_c \oplus \boldsymbol{r}_c) + \boldsymbol{r}_{score}^{q\,\top} (\boldsymbol{e}_c \oplus \boldsymbol{r}_c) \right)$$

Bind the entity and relation vectors in the neighborhood, and then take a **weighted sum** of all of the bindings

$$\mathtt{M}_i = \sum_c \text{weight}(\boldsymbol{e}_c, \boldsymbol{r}_c | \boldsymbol{e}_i, \boldsymbol{r}_q) \mathbb{B}(\boldsymbol{r}_c, \boldsymbol{e}_c)$$

HMem Architecture

Score neighborhood entries → **Compute bindings** → **Sum weighted bindings** → Outputs **MemState**



Compute Weight Matrix (from MemState) → **Optimize rep**

Inference

After optimization, the memory state should include new neighborhood entries that answer the query

We decode these using the corresponding **unbinding function**

$$r_{\text{hyponym}} \cdot \hat{\mathtt{M}}_{\text{canine}}$$ **Dot product (TPR)**

$$r_{\text{hyponym}} \star \hat{\mathtt{M}}_{\text{canine}}$$ **Circular correlation (HRR)**

"Is steppe_wolf a type of canine?"

If yes: $$r_{\text{hyponym}} \cdot \hat{\mathtt{M}}_{\text{canine}} \approx e_{\text{steppe\_wolf}}$$

Results

	WordNet					Freebase				
Model	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
DistMult [Yang et al., 2015]†	457	.790	-	-	.950	**36**	.837	-	-	**.904**
ComplEx [Troullion et al., 2016]	-	.941	.936	.945	.947	-	.692	.599	.759	.840
R-GCN+ [Schlichtkrull et al., 2017]	-	.819	.697	.929	**.964**	-	.696	.601	.760	.842
ConvE [Dettmers et al., 2017a]	374	.943	.935	.946	.956	51	.657	.558	.723	.831
SimplE [Kazemi and Poole, 2018]	-	.942	.939	.944	.947	-	.727	.660	.773	.838
HypER [Balazevic et al., 2019a]	**431**	**.951**	**.947**	**.955**	.958	44	**.790**	**.734**	**.829**	.885
TorusE [Ebisu and Ichise, 2018]	-	.947	.943	.950	.954	-	.733	.674	.771	.832
HMem-CConv	262	.927	.913	.939	.946	**24**	.664	.548	.749	.867
HMem-CConv+	227	.933	.919	**.945**	**.952**	**24**	.664	.547	.749	.866
HMem-CConv$_\infty$	308	.884	.851	.912	.934	39	.488	.363	.554	.734
HMem-CConv$_\infty$+	**183**	.899	.866	.930	.951	39	.481	.357	.546	.725
HMem-CConv$_{im}$	344	**.936**	**.929**	.942	.947	25	**.728**	**.637**	**.795**	**.881**
HMem-TPR	253	.934	.923	.944	.948	30	.590	.478	.660	788
HMem-TPR+	**174**	**.944**	**.932**	**.955**	**.960**	29	.592	.479	.662	.791
HMem-TPR$_\infty$	395	.874	.823	.922	.939	38	.612	.517	.669	.782
HMem-TPR$_\infty$+	323	.879	.24	.930	.950	37	.616	.521	.674	.786
HMem-TPR$_{im}$	245	.936	.924	.947	.952	**24**	**.790**	**.731**	**.831**	**.886**

SOTA

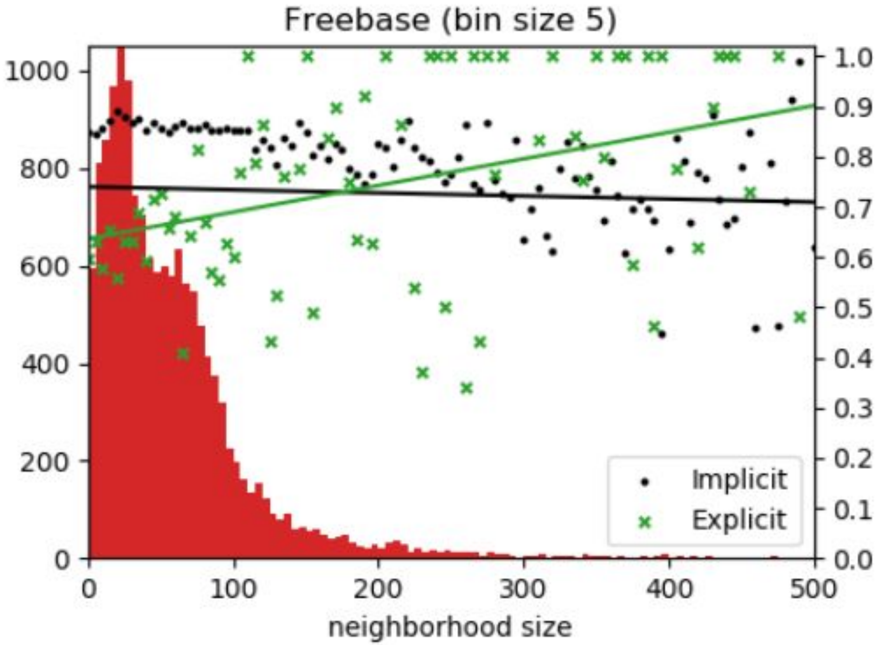
HRR

TPR

Non-compositional (implicit binding) models perform best on Freebase

WordNet: Best Model is TPR with HMax

Implicit vs Explicit Binding



Model	100	200	300	400	500	600
Implicit	.862	.816	.793	.702	.741	.617
Explicit	.632	.746	.772	.856	.835	.900

Implicit > Explicit Binding **only for entities with small neighborhoods**

Why? Embeddings with large neighborhoods have more training instances, but represent more superpositions, meaning more intrusion during unbinding

The **optimal embedding of the memory** is a weighted sum of ALL the neighbor TPRs

$$\mathrm{M}_{\mathrm{cat}} = \sum_{i,j} p(\boldsymbol{r}_i, \boldsymbol{e}_j | \boldsymbol{e}_{\mathrm{cat}}) \ \boldsymbol{r}_i \otimes \boldsymbol{e}_j$$

(learned embeddings)

Scalability considerations

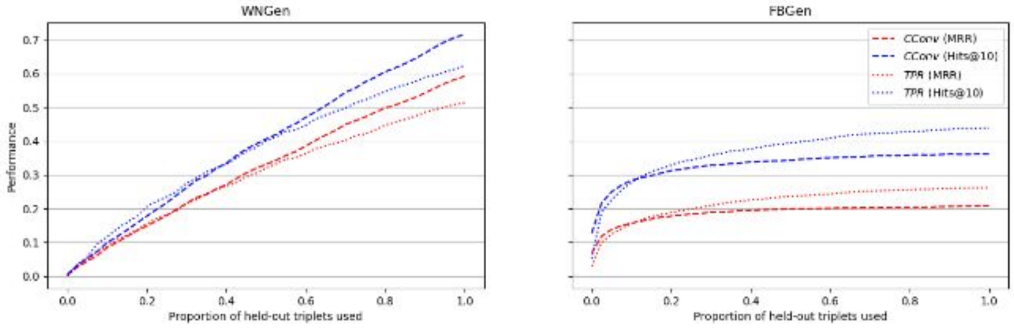
Compositional entity representation allows the model to obtain representations for entities that did not occur in training ⇒ **generalization to novel entities**

2 new datasets: WNGen and FBGen:
Subsets holding out all triplets involving a set of test entities

	heldout	train	valid	test	obs
WNGen	1.5K	141K	1.7K	1.7K	6.8K
FBGen	1K	496K	15K	15K	62K

	Model	MR	MRR	H@1	H@3	H@10
WNGen	CConv	2286	.487	.426	.527	.594
	CConv+	**1359**	**.592**	**.518**	**.647**	**.716**
	TPR_∞	2127	.435	.373	.476	.540
	$\text{TPR}_\infty+$	1507	.514	.448	.565	.624
FBGen	CConv_∞	378	.205	.130	.225	.358
	$\text{CConv}_\infty+$	373	.207	.131	.251	.361
	TPR_∞	401	.252	**.173**	**.299**	**.439**
	$\text{TPR}_\infty+$	**397**	**.263**	**.173**	**.299**	**.439**

Table 5.4: Results on the KBEGen task.



Performance improves smoothly as more triplets are added to the observed subgraph--system extensibility w/out retraining

Spatial Attention Networks

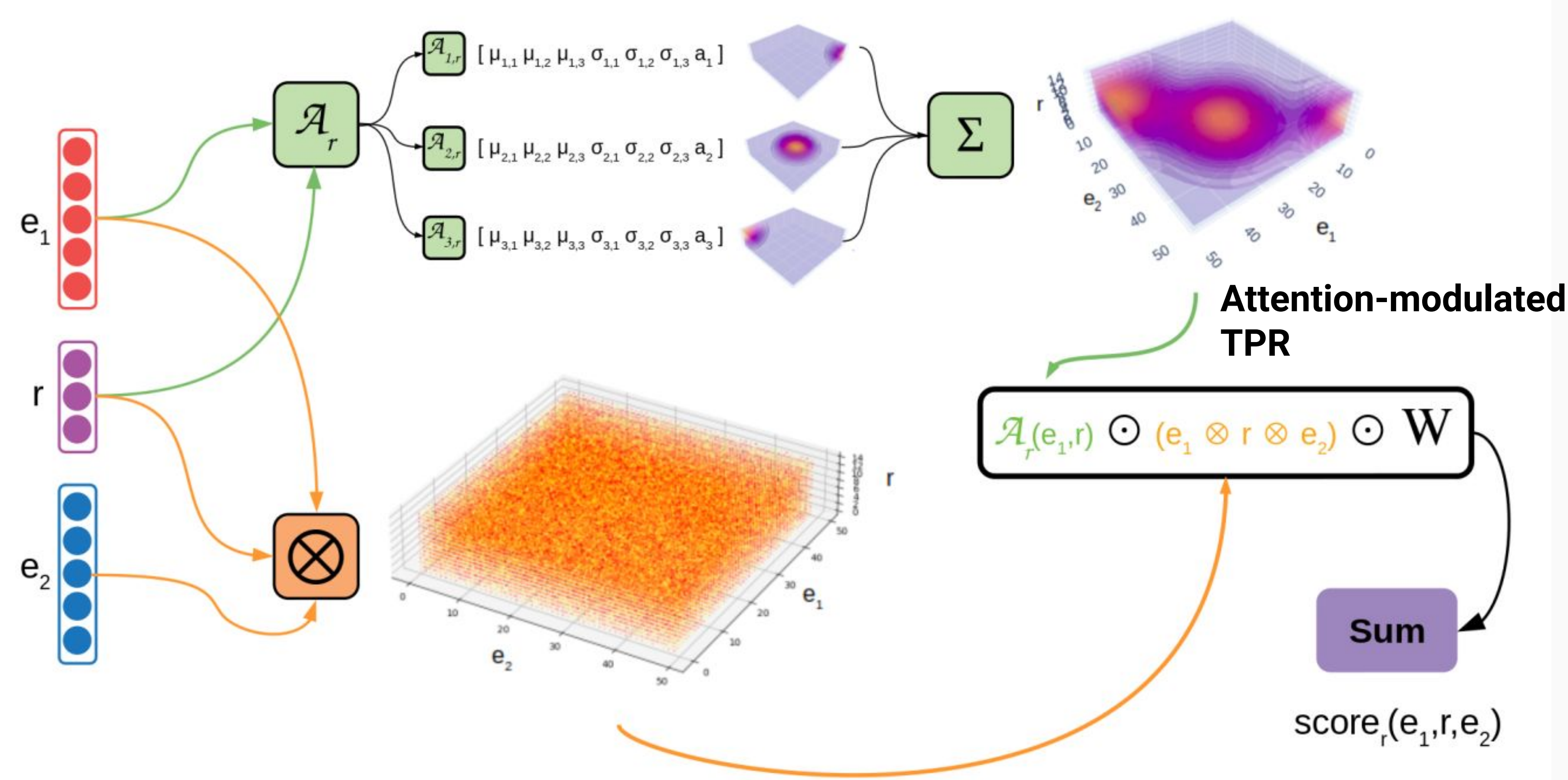
Tensor Product Representations have an implicit spatial structure defined by the coordinates of the involved vectors

SAN input structures: 3-way tensor products of entity and relation vectors

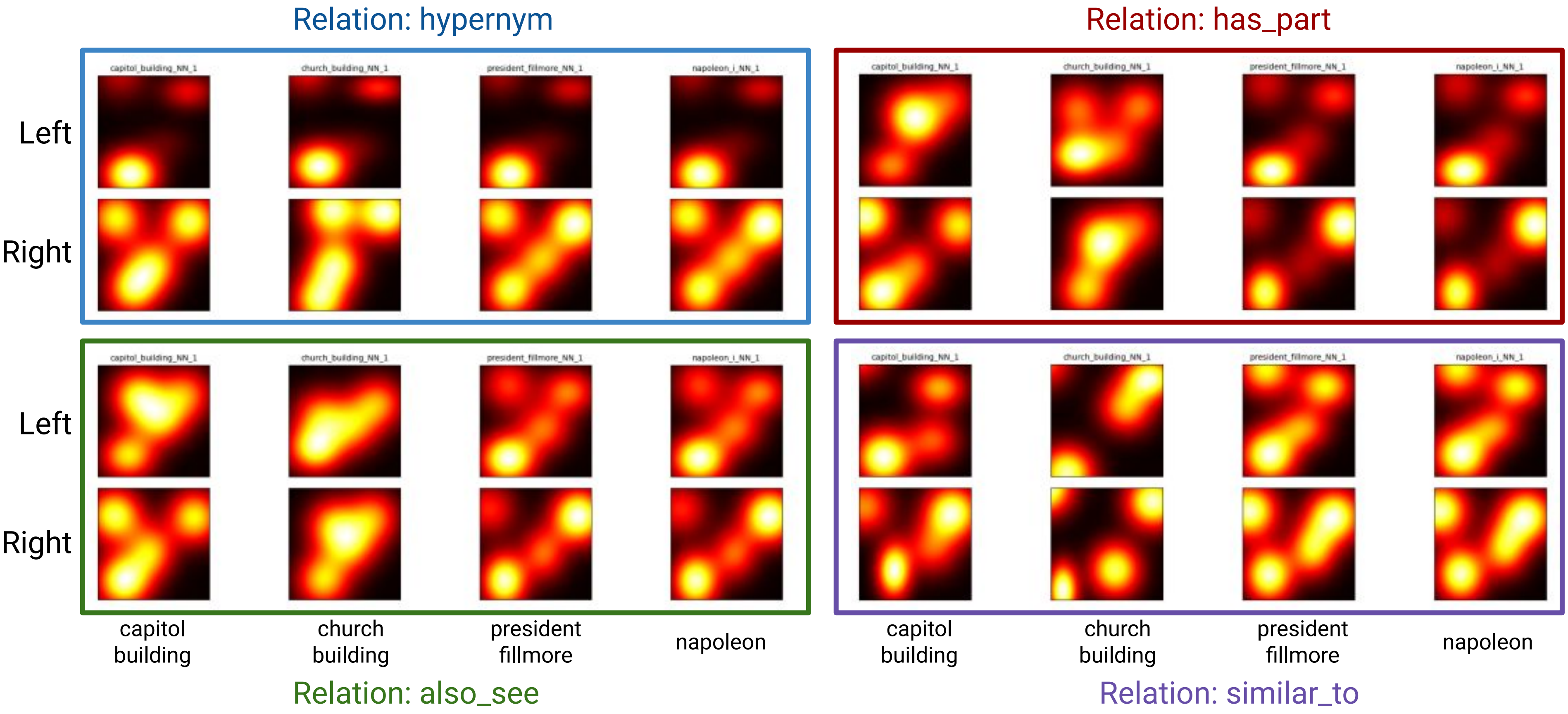
⇒ 3d volumes with 3 spatial coordinates

Can this spatial structure be used as an organizing principle for knowledge representations?

Spatial attention modules: Output attention distributions on the TPR components



TPR module: The triplet representation



Learned Attention Distributions

Results

New dataset assembled from WikiData

Model	MR	MRR	Hits@1	Hits@3	Hits@10
TPR Base	1164	.267	.197	.295	.405
SAN 3H	**983**	**.292**	**.220**	**.326**	**.421**

Table 6.5: Results on the Companies dataset.

SAN outperforms TPR on the **Companies** dataset

WN18RR ("challenge" subset of WordNet)

Model	MR	MRR	Hits@1	Hits@3	Hits@10
M3GM† [Pinter and Eisenstein, 2018]	2193	.498	.454	-	.590
GAAT [Wang et al., 2019]	**1270**	.467	.424	.525	.604
Inverse Model [Dettmers et al., 2017b]	13526	.348	.348	.348	.348
TPR Base	3858	.364	.344	.371	.398
SAN 2H	3463	.376	.353	.386	.416
Inverse Model+rev	13526	.348	.348	.348	.348
TPR Base+rev	**1180**	**.599**	**.572**	**.613**	**.645**
SAN 4H+rev	1656	**.605**	**.580**	**.619**	**.644**

Baseline symbolic model (inverse relations)

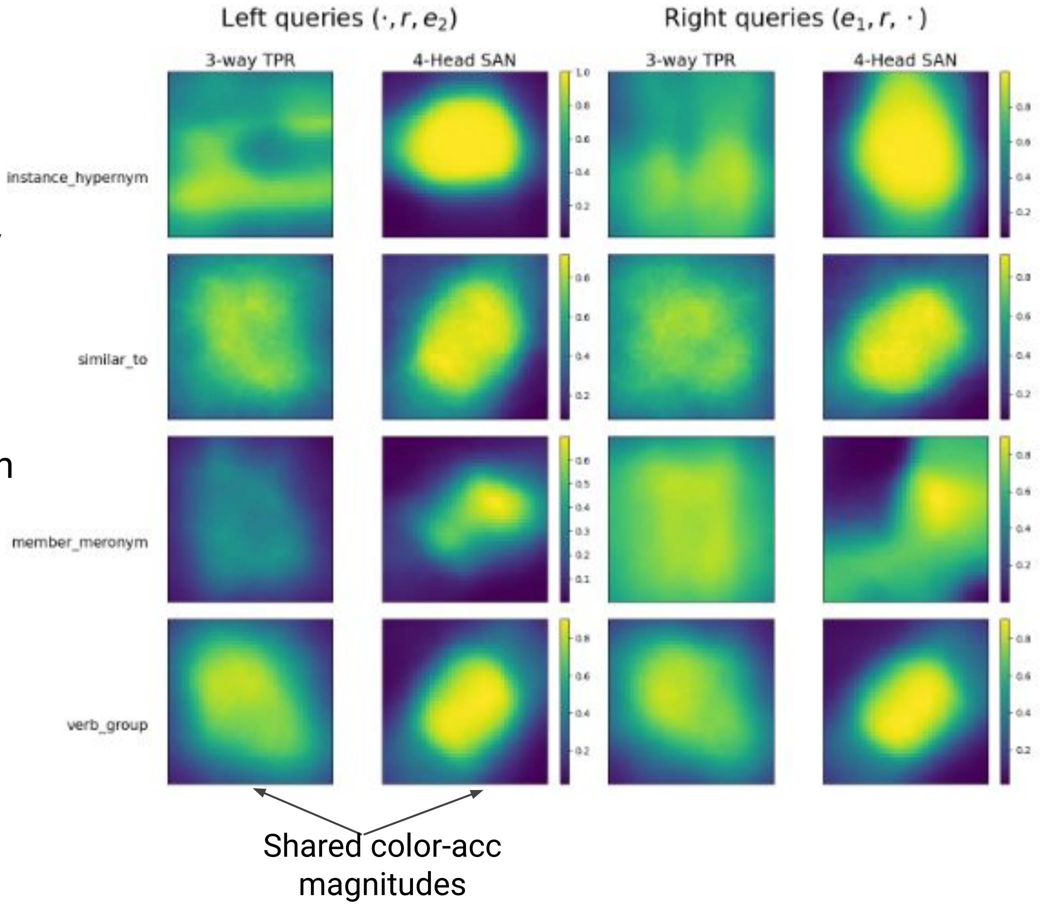
TPR & SAN both outperform the SOTA on WN18RR

Spatial arrangement of features

Accuracy (MRR) when placing the searchlight at each point on the entity1-entity2 grid

3-way TPR: diffuse & lower accuracy distribution (highly distributed representations)

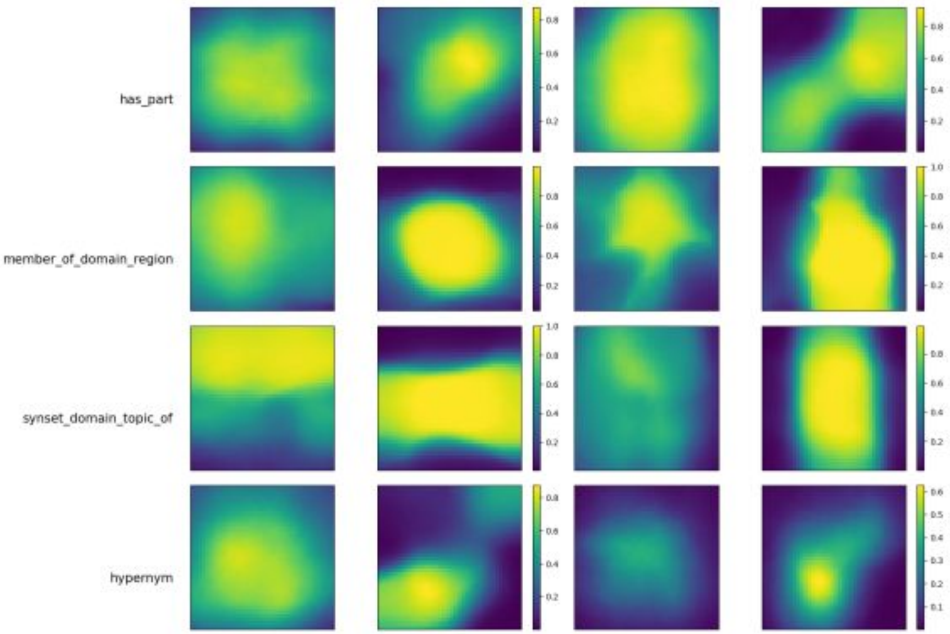
SAN Network: High accuracy in local regions. Relation-specific information tightly localized (semi-localist rep)

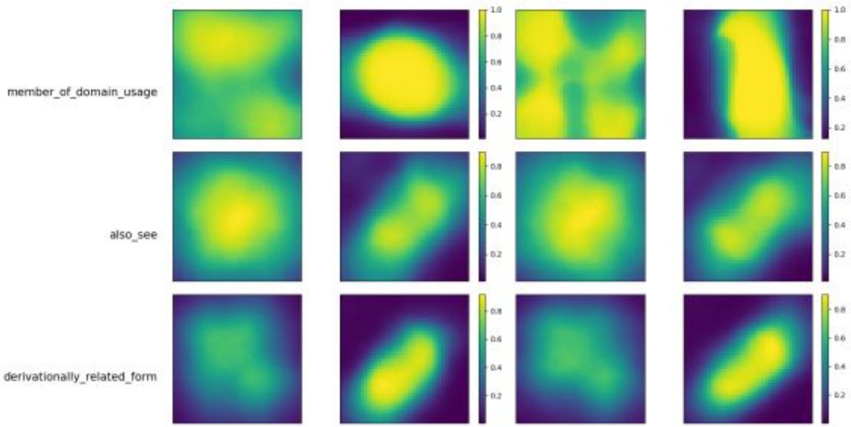


Conclusion

- Explicit binding models provide an implementationalist account of symbol-processing in neural networks (+ similarity & other properties tough to capture in a symbolic model)
- When non-compositional processes come in—e.g. interactive meaning-modulation in coercion/copredication—we can use mechanisms like Harmony Maximization to modulate the representation
- Each of the models presented operates at the SOTA for knowledge base representation
- We hope this work brings attention & interest to classical binding models as candidates for cognitive theories

More searchlights





HRRs & TPRs (the full pipeline)