Acquiring Complex Concepts with Comparative Learning

Diego Calanzone, Filippo Merlo
University of Trento

Presenter: **Diego Calanzone, Filippo Merlo**

Seminar: **Grounded Language Processing**

Academic year: **2023/2024**

Table of Contents

Table of Contents

UNIVERSITÀ DI TRENTO | CiMeC Center for Mind/Brain Sciences

The research question

Machines don’t learn efficiently.

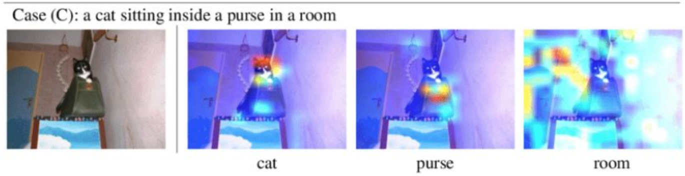
- Humans learn <u>continually</u>, machines <u>once for all</u>.
- Humans learn from <u>few examples</u>, machines need to <u>iterate</u> through <u>many of them</u>.

	Accuracy	Majority Vote on Full Dataset
Zero-shot human	$\boxed{53.7}$	57.0
Zero-shot CLIP	**93.5**	**93.5**
One-shot human	$\boxed{75.7}$	80.3
Two-shot human	75.7	85.0

Learning Transferable Visual Models From Natural Language Supervision. Radford et al. 2021

How do we acquire new words/symbols?

- **What to learn?** By **extracting** key information of a concept to re-use it in the future. [Tomasello and Farrar, 1986]
- **How to learn?** By finding commonalities and differences between two pieces of information. [Gentner and Markman, 1994]



Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. Huang et al. 2020

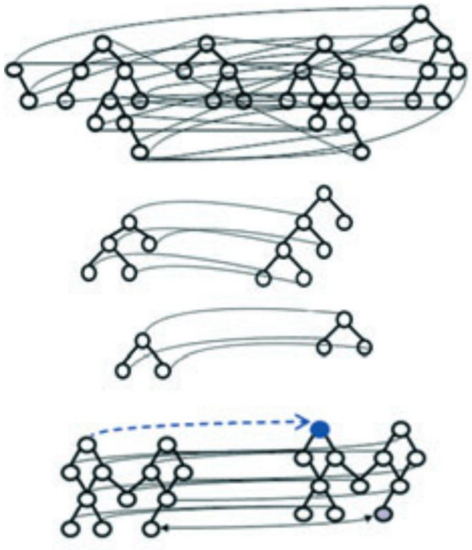
Background theories

The theory of Structure-Mapping

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. Cognitive science, 7(2):155–170.

Comparison allows **to attend relational structures** in inputs, highlight differences.
Human infants learn this with few examples.

SME (the Structure-mapping Engine)



Three Stages

1. Local identity matches made in parallel
2. Structural consistency enforced->small submappings (*kernels*)
3. Kernels combined into maximal structurally consistent mapping
   - Structural evaluation
   - Candidate inferences
   - Alignable differences

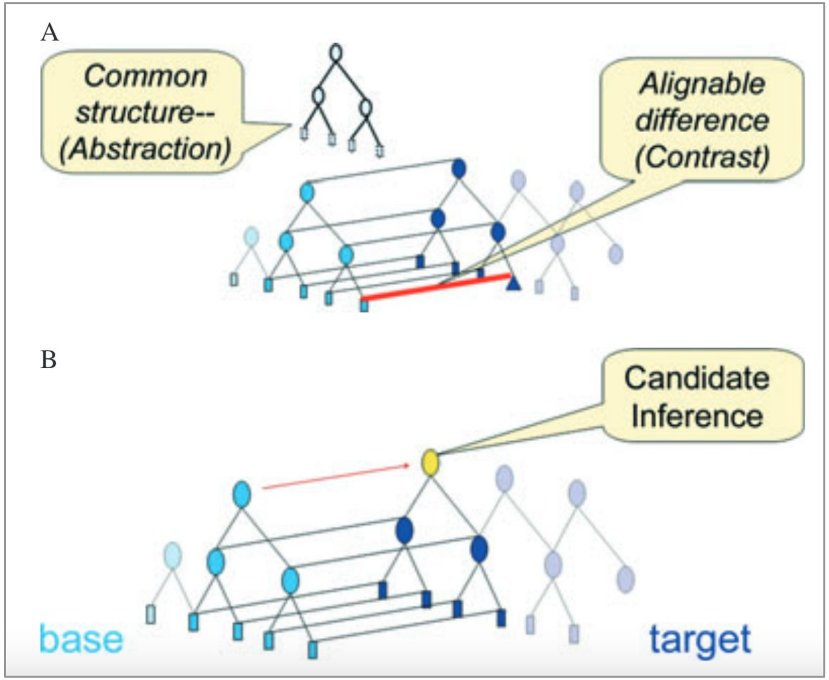
Gentner D. Bootstrapping the mind: analogical processes and symbol systems. Cogn Sci. 2010 Jul;34(5):752-75. doi: 10.1111/j.1551-6709.2010.01114.x. PMID: 21564235.

Background theories

The theory of Structure-Mapping

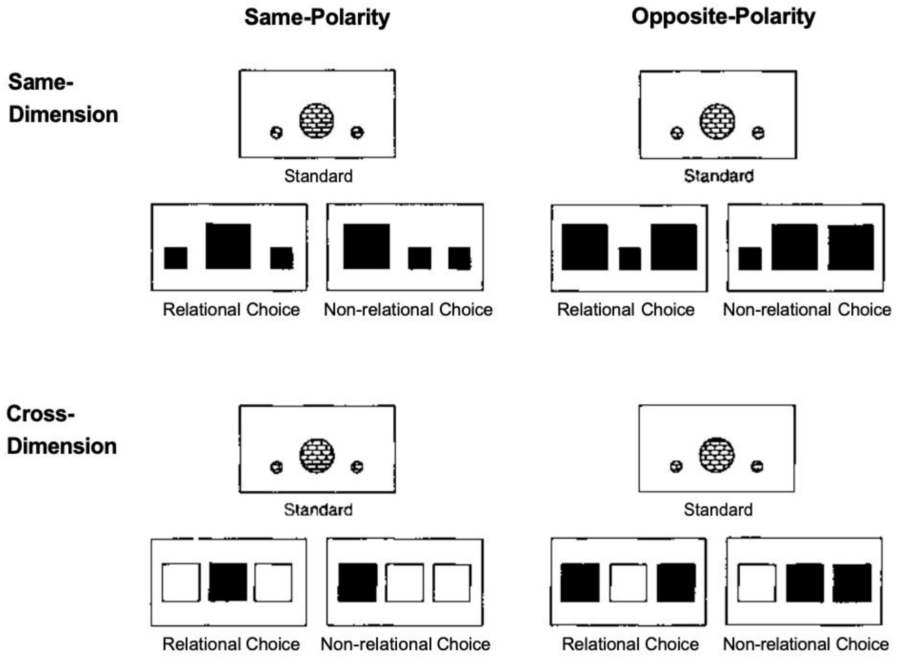
Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. Cognitive science, 7(2):155–170.

Comparison allows **to attend relational structures** in inputs, highlight differences.
Human infants learn this with few examples.



[Gentner D. Bootstrapping the mind: analogical processes and symbol systems. Cogn Sci. 2010 Jul;34(5):752-75. doi: 10.1111/j.1551-6709.2010.01114.x. PMID: 21564235.]

Background theories



Progressive Alignment

Hespos et al. 2020 Structure-mapping processes enable infants' learning across domains including language

Kotovsky, Gentner. 1996. Comparison and categorization in the development of relational similarity.

Aligning highly similar inputs invite young children to **reason about relational structures** for abstraction and characteristic learning.

Related work

Learning in neural networks

- **Pre-training**: using internet-scale collections of data and hoping to well transfer knowledge

- **Continual learning**: training the model on subsequent sets of knowledge without forgetting
  - Rehearsal/replay methods
  - Learn with penalties on forgetting
  - Domain adaptation/generalization

→ What about gradually learning with developmental psychology theories in mind?

Table of Contents

Proposed Methodology: Base Work

[Human Inspired Progressive Alignment and Comparative Learning for Grounded Word Acquisition. Bao et al. 2023]

Human Inspired Progressive Alignment and Comparative Learning for Grounded Word Acquisition

Yuwei Bao[†] **Barrett Martin Lattimer**[§*] **Joyce Chai**[†]
[†]Computer Science and Engineering, University of Michigan [§]ASAPP
{yuweibao, lattimer, chaijy}@umich.edu

- A curated **dataset**
  - **S**imulated **O**bjects for **L**anguage **A**cquisition (**SOLA**)
- A **methodology** for:
  - Grounded word acquisition (Comparative Learning)
  - Concept learning: filtration & representation mapping
- **Benchmarks** to test:
  - Multi-attribute recognition
  - Continual learning
  - Novel composition reasoning

Proposed Methodology: Dataset

Human Inspired Progressive Alignment and Comparative Learning for Grounded Word Acquisition. Bao et al. 2023

Low noise and **distinct attributes**

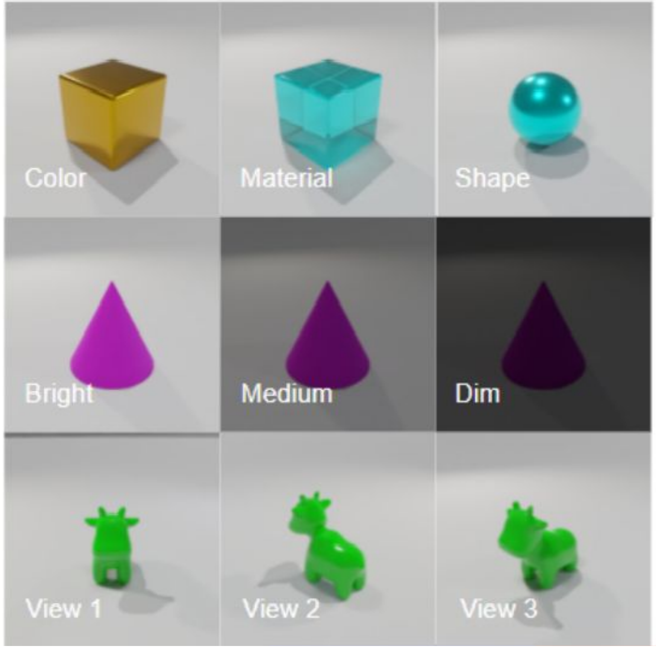
Facilitates efficient sample comparisons and mapping of language features to grounded concepts.

Combination of 8 colors, 11 shapes, and 4 materials

Variations in 3 light settings and 6 camera angles.

6336 RGBA images of synthetic objects.

Total of **23 concepts**



SOLA dataset renders. Bao et al. 2023

Proposed Methodology: Dataset

Human Inspired Progressive Alignment and Comparative Learning for Grounded Word Acquisition. Bao et al. 2023

Variation Test set (Dtest_v) → # 989

(To assess model generalizability and robustness)

- stretching along the x, y, and z axes
- shade changes
- size alterations to medium or small dimensions

A Novel Composition Test set (Dtest_nc) → # 1242

(To evaluate the novel composition capability of the methods)

- 9 exclusive learning attribute pairs

Train set (Dtrain) → # 5094

- Remaining pairs

SOLA dataset renders. Bao et al. 2023

Proposed Methodology: Dataset

Acquiring Complex Concepts with Comparative Learning, Calanzone and Merlo 2024

(For complex concept acquisition)

Dtrain_complex → 5760

- Merging the Dtrain and Dtest_nc datasets

Dtest_no → 576

- We subtract a set of 32 objects from Dtrain_complex to create a test dataset of objects unseen during training

(new runs are: Dtrain_complex: 5076 and Dtest_no: 1260)

Split	Num Objects
D_{train}	5094
$D_{train\_complex}$	5760
D_{test_nc}	1242
D_{test_no}	576
D_{test_v}	989

Table of Contents

Proposed Methodology: Comparative Learning

- For each learned concept l_i in an unconstrained set $L = \{l_1, l_2, \ldots\}$, must be assembled a batch of samples $\mathcal{B}_s = \{a_1^{l_i}, \ldots, a_n^{l_i}\}$, that share the label l_i for similarity learning, and a batch of samples $\mathcal{B}_d = \{b_1^{l_j}, \ldots, b_n^{l_j}\}, j \neq i$ that cannot be described by l_i for difference learning.

- The process of SIM_{l_i} (1) finds the similarities among the examples in $\mathcal{B}_s$, and extract out the representation REP_{l_i} expected to refer to l_i.
- The process of DIFF_{l_i} (2) highlights the differences between l_i and other non-compatible labels refining the representation REP_{l_i}.

$$\text{REP}_{l_i} = \text{SIM}_{l_i}(\{a^{l_i} \in \mathcal{B}_s\}) \tag{1}$$

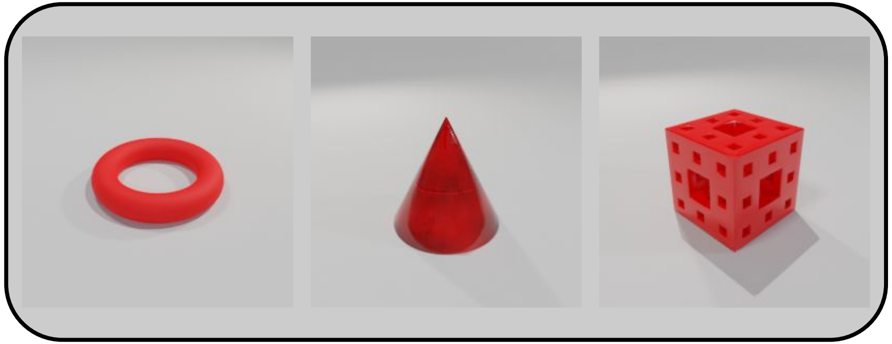
$$\text{REP}_{l_i} = \text{DIFF}_{l_i}(a^{l_i}, \{b^{l_j} \in \mathcal{B}_d\}) \tag{2}$$

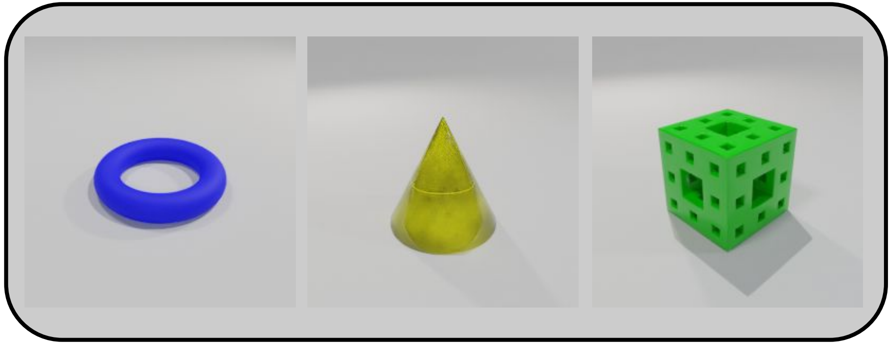
Proposed Methodology: Baseline

For each concept (e.g., "red"):

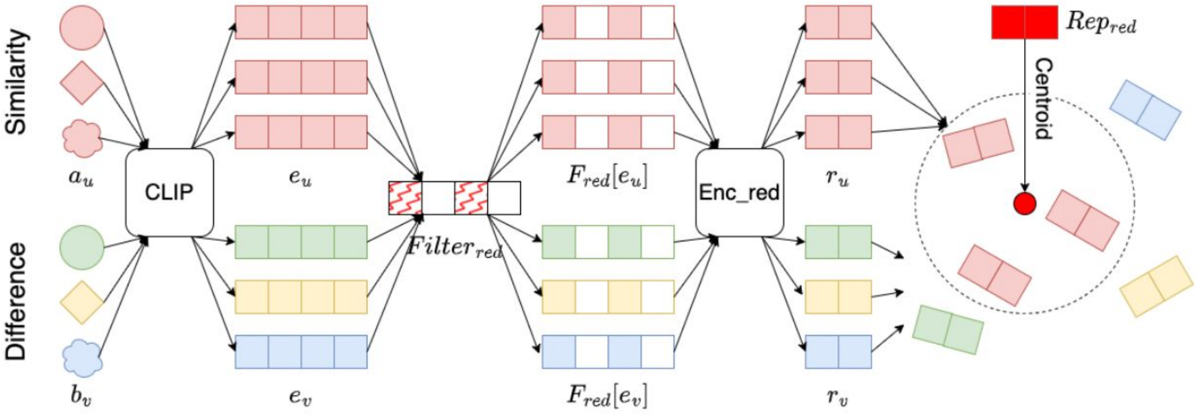
- **Similarity training**: They gather a batch of images of the specified concept ("red").
- **Difference refinement**: Another batch consists of images that are of any other color but still "red" (non-compatible).

They ensure that the rest of the attributes remain consistent across batches for better structural alignment.





Proposed Methodology: Baseline



(a) Filter, Encoder, Representation Learning

Algorithm 1: Comparative Learning-Word l_i

for *Sim and Diff data batches:* $\{\mathcal{B}_s, \mathcal{B}_d\}$ **do**

    // Similarity Learning

    for $a_u \in \mathcal{B}_s$ **do**

        $e_u = \texttt{CLIP\_emb}[a_u]$

        $r_u = \texttt{Enc}_{l_i}[\texttt{F}_{l_i}(e_u)]$

    $\text{Rep}_{l_i} = \texttt{Centroid}[\{r_u\}_{u \in \{1,\cdots,n\}}]$

    // Difference Learning

    for $b_v \in \mathcal{B}_d$ **do**

        $e_v = \texttt{CLIP\_emb}[b_v]$

        $r_v = \texttt{Enc}_{l_i}[\texttt{F}_{l_i}(e_v)]$

    // Loss

    $\text{loss}_s = \sum_u \texttt{Dist}[r_u, \text{Rep}_{l_i}]$

    $\text{loss}_d = \sum_v \texttt{Dist}[r_v, \text{Rep}_{l_i}]$

    $\text{loss} = (\text{loss}_s)^2 + (1 - \text{loss}_d)^2$

    Backpropagate and Optimize

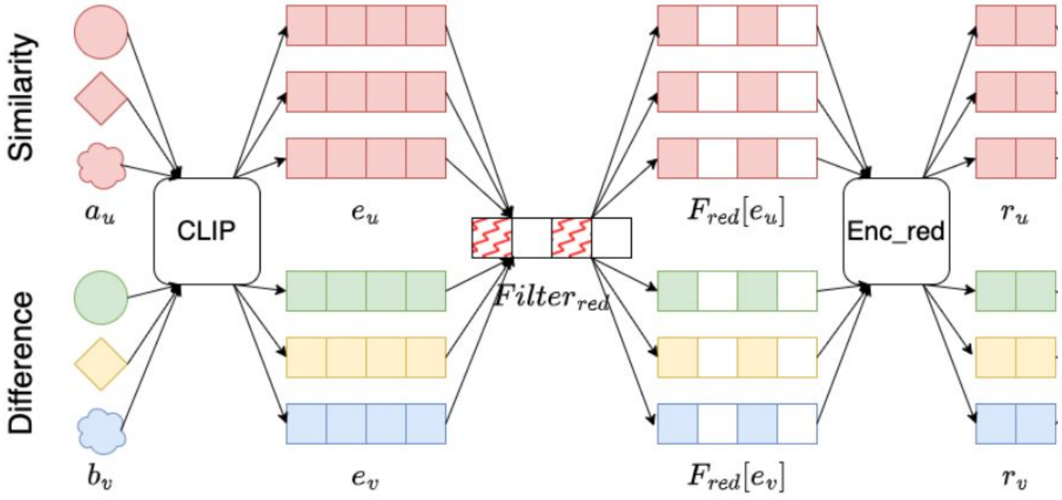
Output: $\{l_i$: $[\text{F}_{l_i}, \text{Enc}_{l_i}, \text{Rep}_{l_i}]\}$

Proposed Methodology: Baseline

Initially, a [pre-trained frozen CLIP image embedding] is utilized.

Image embeddings undergo two processes:

- **Information denoising**: Each image embedding is subjected to an elementwise product with a filter.
- **Attention establishment**: The masked embedding passes through two fully connected layers of an encoder to output a condensed representation.



(a) Filter, Encoder, Representation Learning

Proposed Methodology: Baseline

A **centroid** is computed for the representations of the similarity batch.

The loss function serves two purposes:

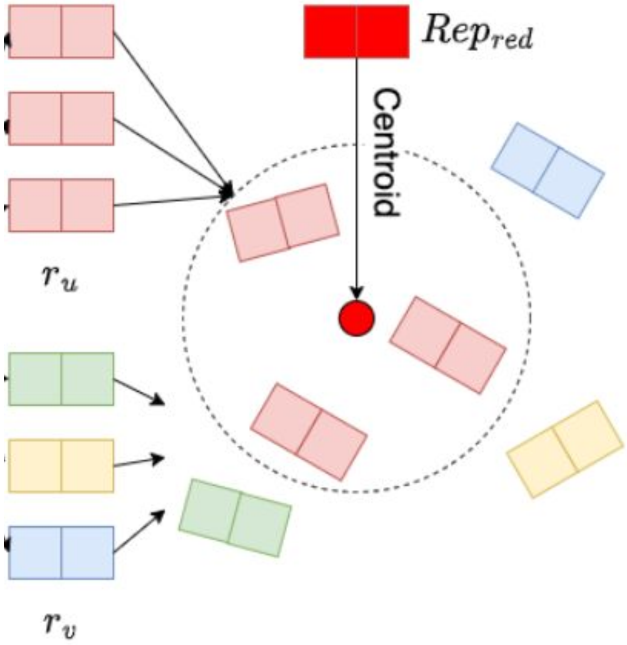
- It **drives the similarity** batch sample representations closer to the centroid.
- It **pushes the difference** batch sample representations further away from the centroid.

// Loss

$$\text{loss}_s = \textstyle\sum_u \texttt{Dist}[r_u, \text{Rep}_{l_i}]$$

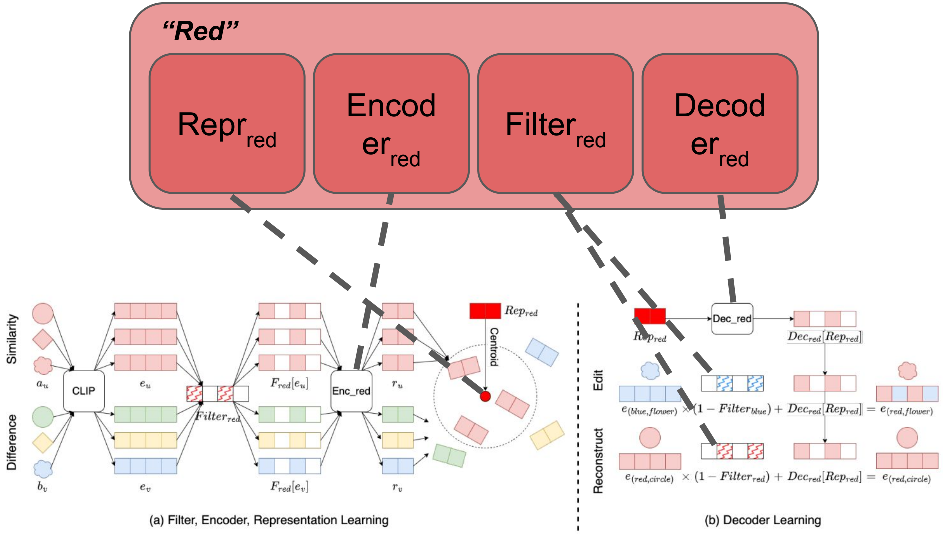
$$\text{loss}_d = \textstyle\sum_v \texttt{Dist}[r_v, \text{Rep}_{l_i}]$$

$$\text{loss} = (\text{loss}_s)^2 + (1 - \text{loss}_d)^2$$



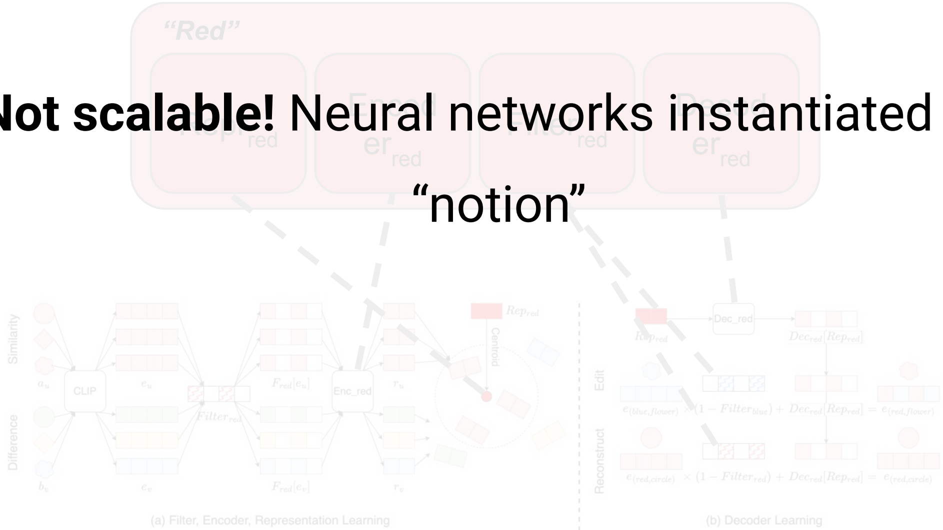
Proposed Methodology: Baseline

- This approach jointly trains the **filter**, the **encoder**, and the **representation**, producing a different set of these three objects **for each of the learned concepts**.



Proposed Methodology: Baseline

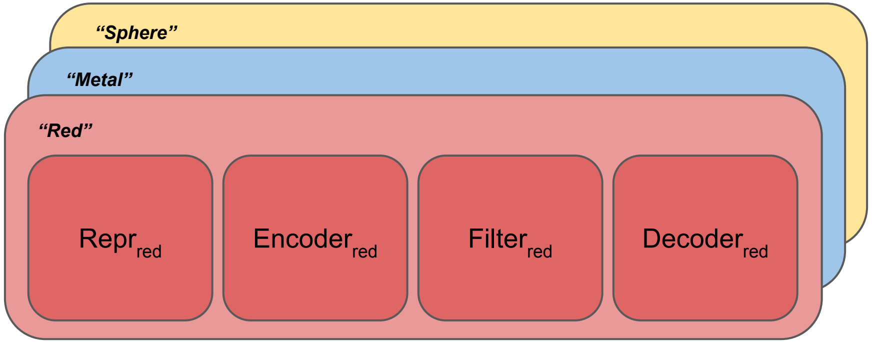
- This approach jointly trains the **filter**, the **encoder**, and the **representation**, producing a different set of these three objects **for each of the learned concepts**.



Not scalable! Neural networks instantiated per "notion"

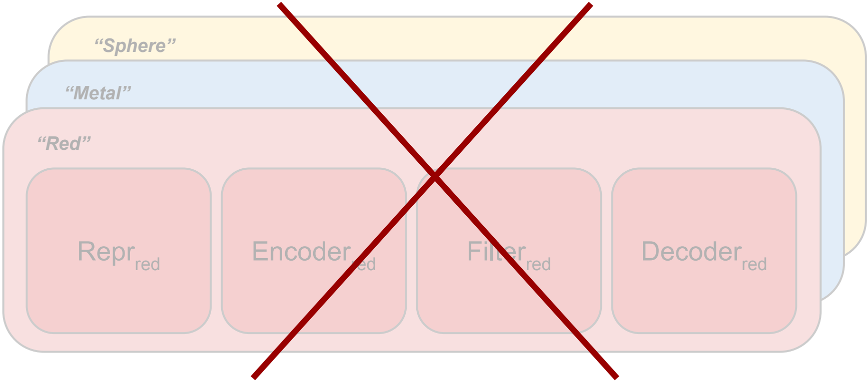
Proposed Methodology: Baseline

- Our first goal is **to unify** the learning of multiple concepts **under the same single architecture** while keeping the same training process.



Proposed Methodology: Baseline

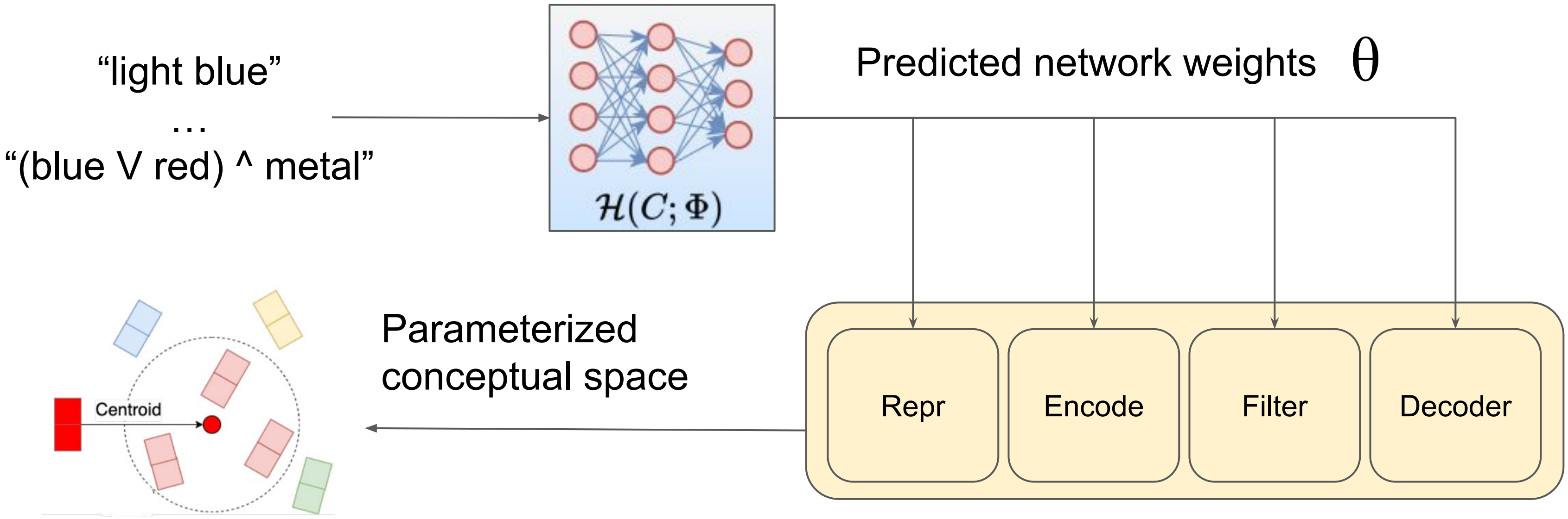
- Our first goal is **to unify** the learning of multiple concepts **under the same single architecture** while keeping the same training process.



Proposed Methodology: Hypernetworks (HyperMem)

Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks. Mahabadi et al. 2021

- Our first goal is **to unify** the learning of multiple concepts **under the same single architecture** while keeping the same training process.



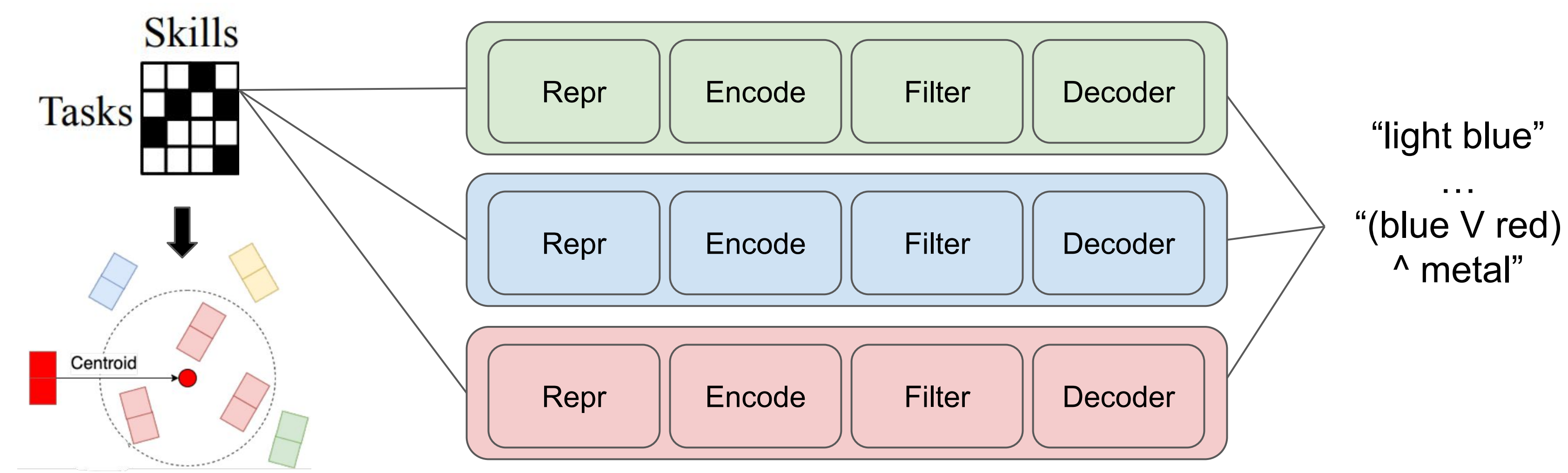
Proposed Methodology: Modular Shared Skills (Polytropon)

Combining Modular Skills in Multitask Learning. Ponti et al. 2023

- Our first goal is **to unify** the learning of multiple concepts **under the same single architecture** while keeping the same training process.

Learned skills allocation

Redundant networks (skills)



UNIVERSITÀ DI TRENTO | CIMeC Center for Mind/Brain Sciences

Proposed Methodology: Modular Shared Skills (Polytropon)

Combining Modular Skills in Multitask Learning. Ponti et al. 2023

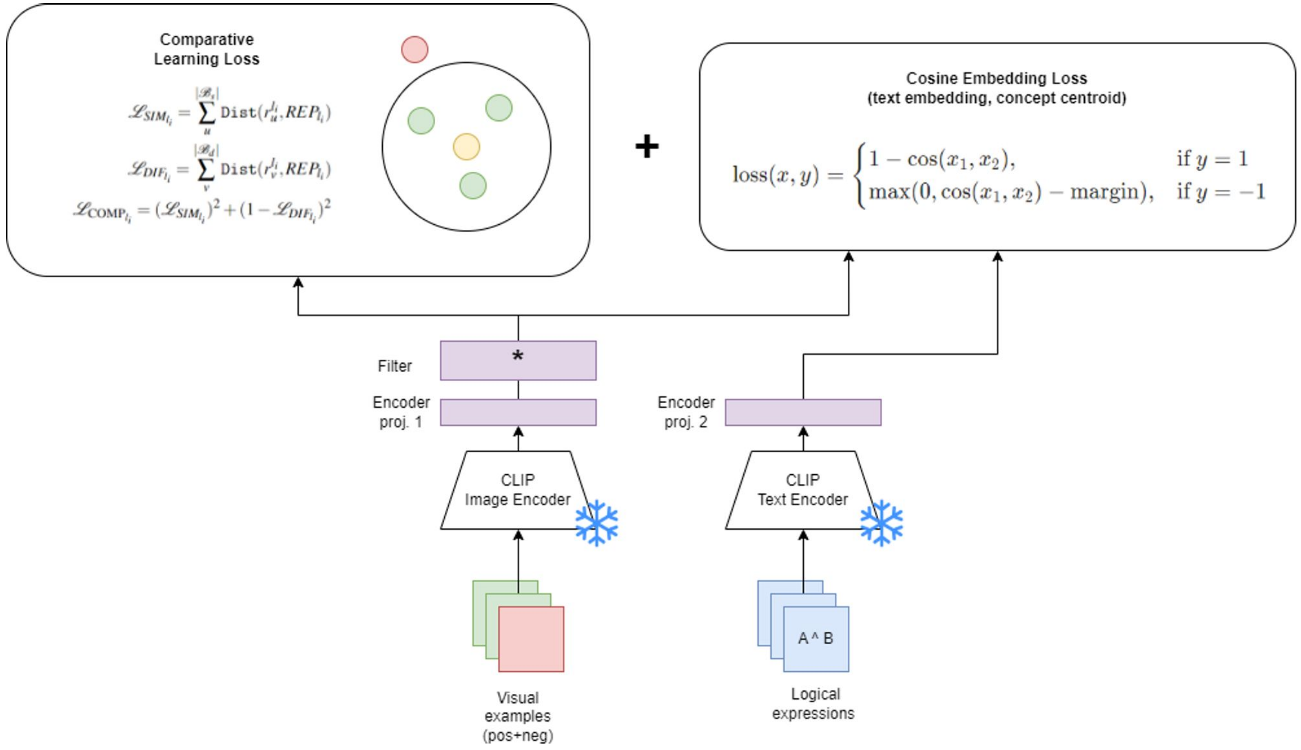


Table of Contents

Proposed Methodology: Compound Logical Concepts

It is possible to teach more complex concepts through the same training process?

We compose simple concepts into logical expressions with basic logic operators: **NOT**, **AND**, and **OR**.

Given an unconstrained set of base concepts (e.g., “red” and “cone”), we considered all possible logical pairs obtaining the set of complex concepts:

- “NOT red”,
- “NOT cone”,
- “red AND cone”,
- “red OR cone”.

Total of **351 new concepts**

Proposed Methodology: Compound Logical Concepts

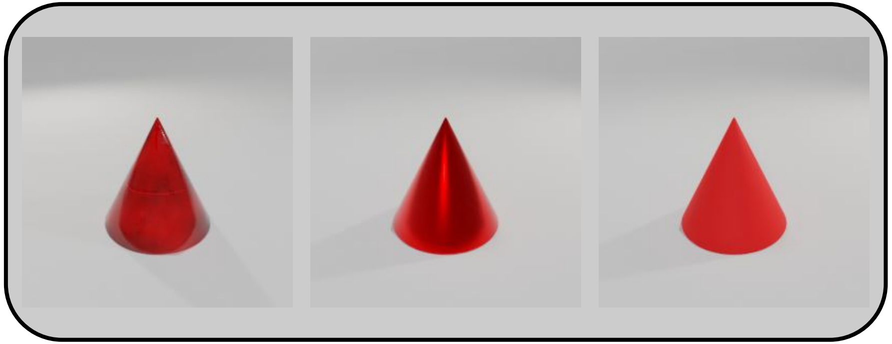
- **Similarity batch** of images with positive samples where the logical relation between the two simple concepts is respected.
- **Difference batch** with negative samples was generated where the relation is violated.

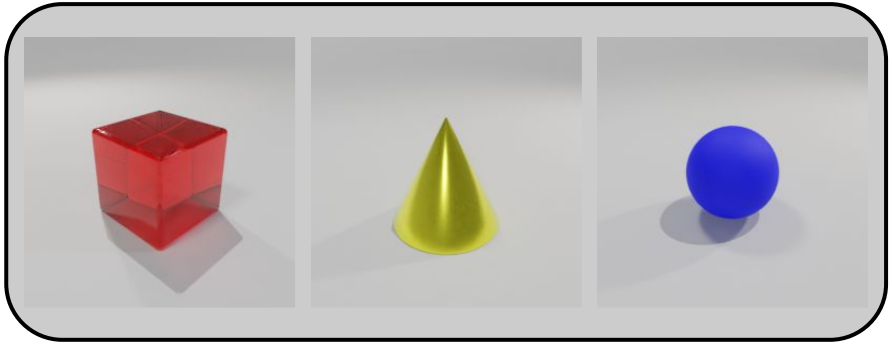
The samples were paired so that, except for the attributes significant for the truth value of the relation, all other features were kept constant.

Proposed Methodology: Compound Logical Concepts

AND

$$\mathscr{B}_s = \{a \,|\, \text{red AND cone}\}$$
$$\mathscr{B}_d = \{a \,|\, \text{red AND NOT cone} \oplus \text{NOT red AND cone} \oplus \text{NOT red AND NOT cone}\} \tag{9}$$

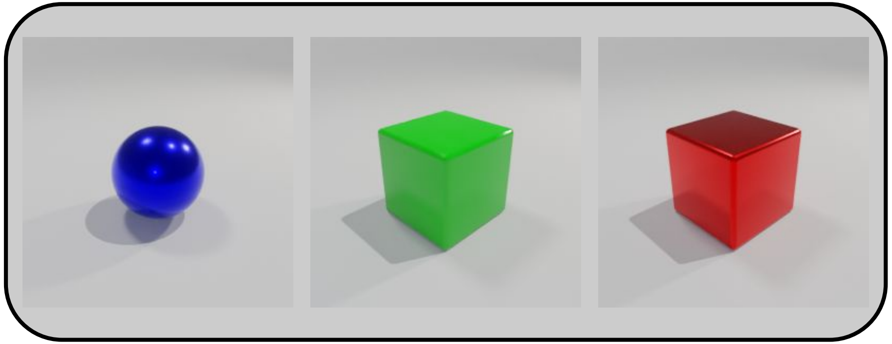




Proposed Methodology: Compound Logical Concepts

OR

$$\begin{aligned} \mathscr{B}_s &= \{a \,|\, \text{metallic AND NOT cube} \oplus \text{NOT metallic AND cube} \oplus \text{metallic AND cube}\} \\ \mathscr{B}_d &= \{a \,|\, \text{NOT metallic AND NOT cube}\} \end{aligned} \tag{10}$$



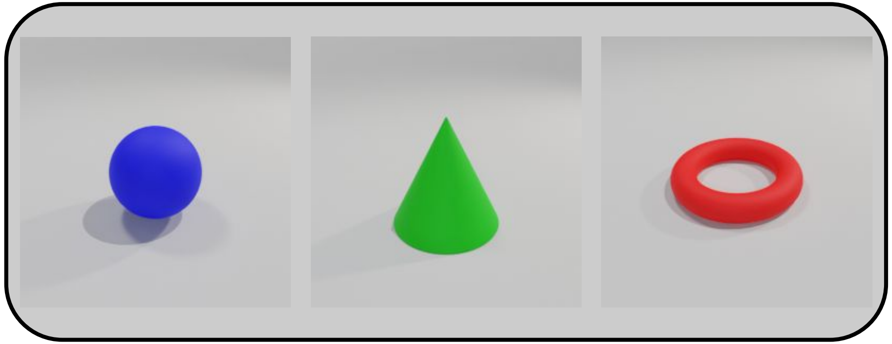


Table of Contents

Evaluation

To test the **acquisition of primitives**, we employ the same cognitive task introduced by Bao et al.: Multi-Attribute Recognition (MAR).

We thus compare:

- the memory-of-networks model (**Baseline**)
- our multi-task hyper-network (**HyperMem**)
- shared modular skills (**Polytropon**)

For **complex logical expressions**, we modify MAR and thus define Logical Pattern Recognition (LPR).

Evaluation: Multi-Attribute Recognition

- Go through the memory
- Apply the corresponding filter and encoder of each concept
- Retrieve the **top-3 concepts** with the least MSE between *Enc_word* and *Learned_rep*

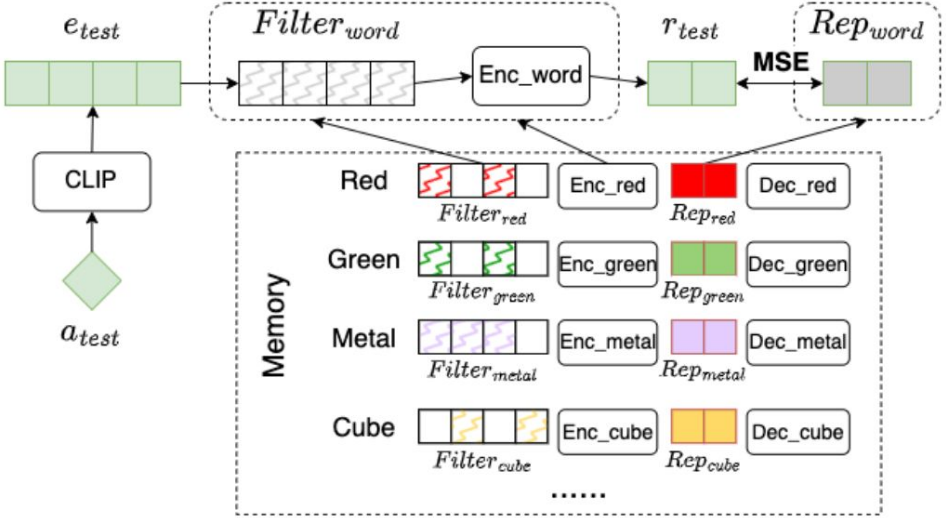


Figure 3: Multi-Attribute Recognition Inference

Evaluation: Multi-Attribute Recognition

Observations:

- Polytropon is comparable to HyperMem, <u>with 10x less parameters</u>
- Distinguishing <u>materials</u> remains the <u>hardest task</u>
- The “baseline” (upper bound) follows a similar pattern

Split	Model	Color	Material	Shape
D_{test_v}	Baseline	0.95	0.75	0.89
	HyperMem	0.56	0.26	0.66
	HyperMem (DER++)	0.74	0.37	0.70
D_{test_nc}	Baseline	0.96	0.48	0.98
	HyperMem	0.37	0.25	0.73
	HyperMem (DER++)	0.71	0.28	0.89
	Polytropon	0.73	0.21	0.67

Multi-Attribute Recognition. Accuracy scores on test variation and novel composition sets. Calanzone and Merlo 2024

Evaluation: Logical Pattern Recognition

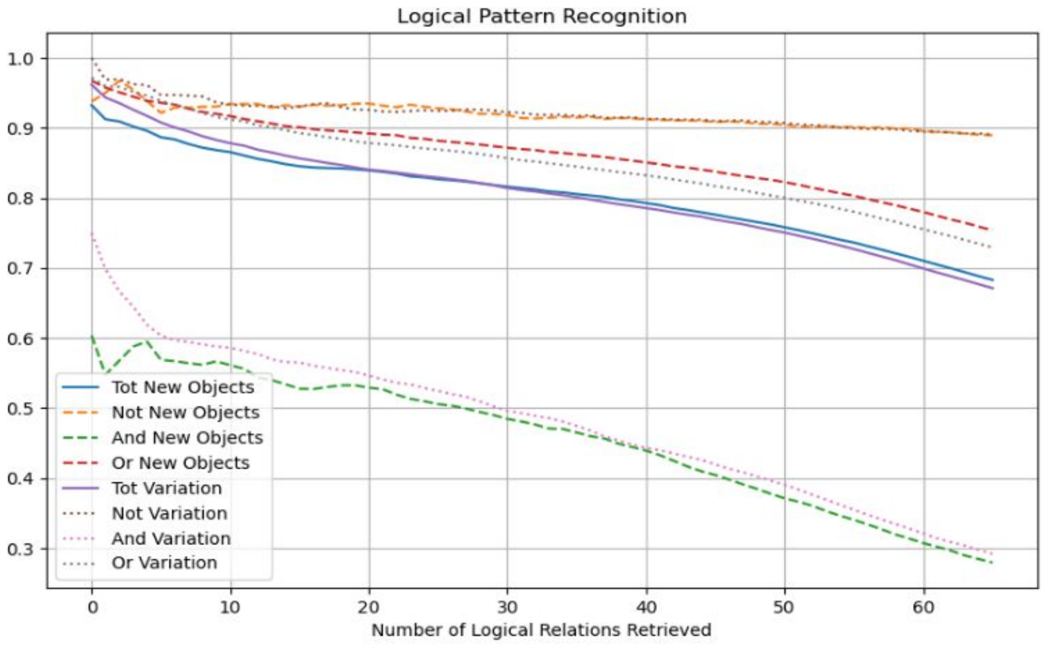
- AND, OR, and NOT relations of the three attributes constituting the image, amount to a total of **66 true relations per image**.
- Within the **top-66 concepts** retrieved, we count as hit only the ones that are true for the evaluated image.

Purple Plastic Torus

1. Aqua or Torus,
2. Red or Torus,
3. Purple or Glass,
4. Brown or Torus,
5. Purple and Plastic,
6. Plastic and Torus,
7. Not Plastic,
8. Purple or Gear,
9. Purple or Rubber,
10. Purple and Torus

Evaluation: Logical Pattern Recognition

- LPR in multiple iterations, **systematically altering the top-k parameter** for concept retrieval, ranging from 1 to 66
- AND relations presents a worse performance



Evaluation: Logical Pattern Recognition

Observations:

- We only test the baseline model.
- Logical conjunction (AND) is a stricter operator, learned with more difficulty
- Early experiments with HyperMem and Polytropon haven't converged or shown comparable results

Top-k Num	Split	Tot	NOT	AND	OR
10	D_{test_no}	0.8682	0.9305	0.5667	0.9201
	D_{test_v}	0.8827	0.9364	0.5878	0.9158
20	D_{test_no}	0.8411	0.9344	0.5326	0.8935
	D_{test_v}	0.8435	0.9261	0.5516	0.8809
30	D_{test_no}	0.8185	0.9201	0.4900	0.8738
	D_{test_v}	0.8179	0.9248	0.5025	0.8599
40	D_{test_no}	0.7957	0.9137	0.4444	0.8533
	D_{test_v}	0.7884	0.9150	0.4478	0.8348
50	D_{test_no}	0.7621	0.9057	0.3776	0.8259
	D_{test_v}	0.7543	0.9081	0.3965	0.8035
60	D_{test_no}	0.7157	0.8990	0.3136	0.7847
	D_{test_v}	0.7052	0.8960	0.3276	0.7605
66	D_{test_no}	0.6830	0.8893	0.2794	0.7538
	D_{test_v}	0.6712	0.8906	0.2919	0.7294

Top-K accuracy in Logical Pattern Recognition. Calanzone and Merlo 2024

Evaluation: Logical Pattern Recognition

- We **analyzed the error rate contributions of the three distinct categories of AND patterns**: Color AND Material, Color AND Shape, and Material AND Shape.

- The error contribution from **Color AND Shape** remained <u>negligible</u> in the initial trials but exhibited an upward trend as the number of representations retrieved increased.

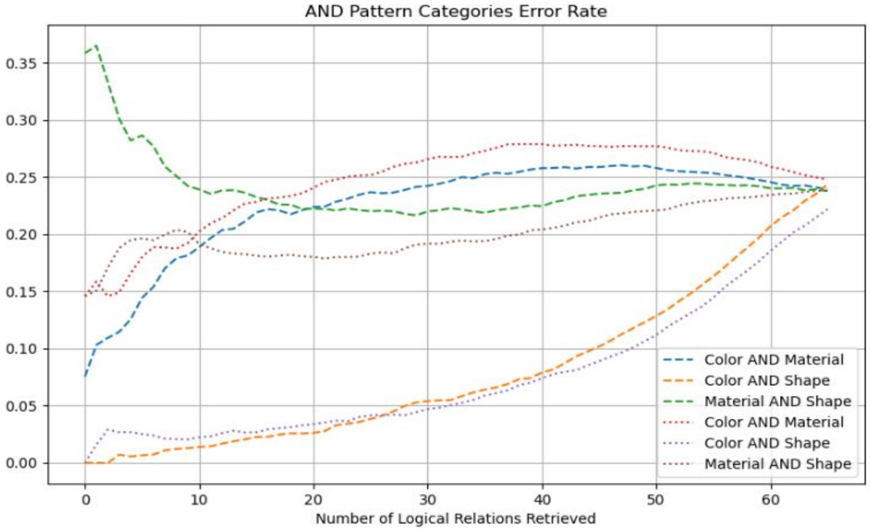
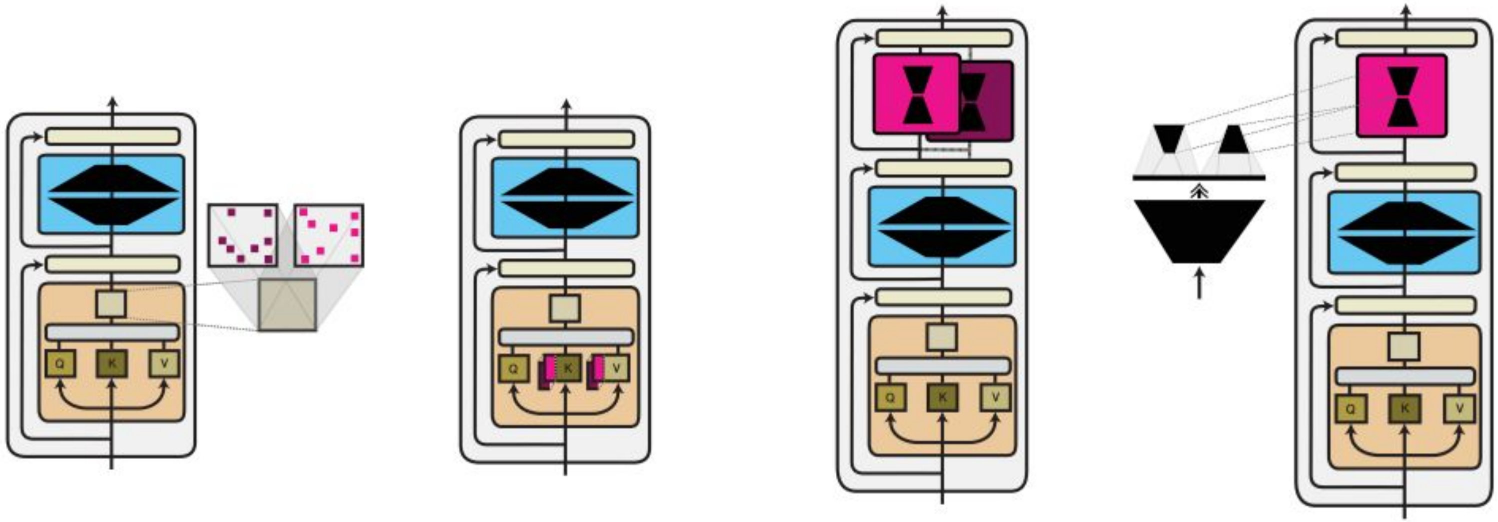


Table of Contents

Where to go from here?

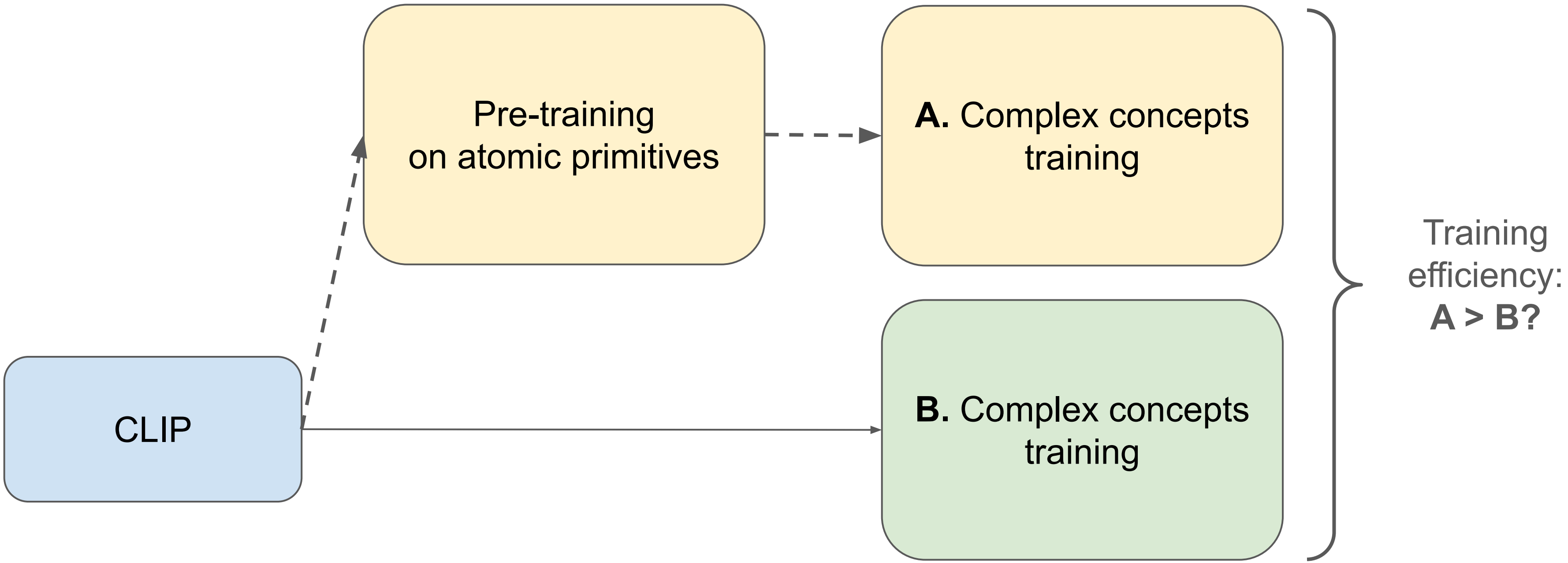
- [Modular Deep Learning (Ponti et al. 2023)] suggests sound and efficient multi-task learning architectures.



(a) Parameter Composition (b) Input Composition (c) Function Composition (d) Hypernetwork

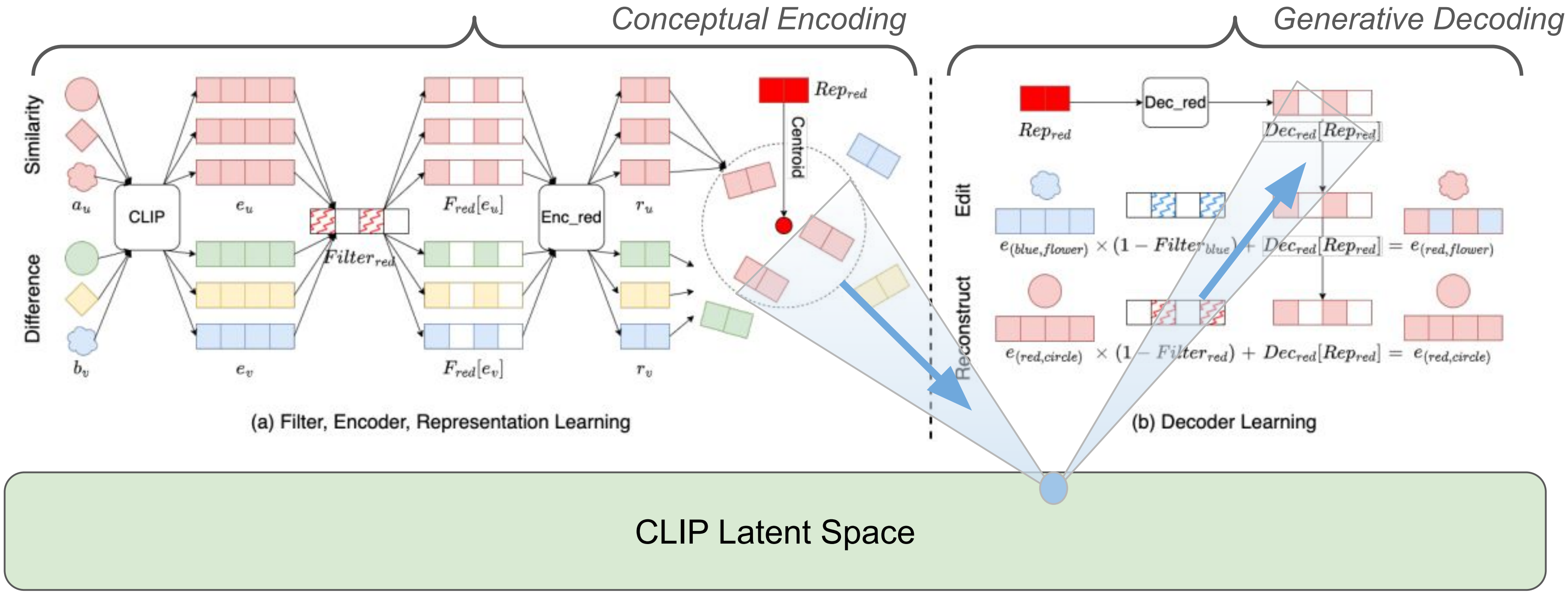
Where to go from here?

- How to test the effects of progressive alignment?



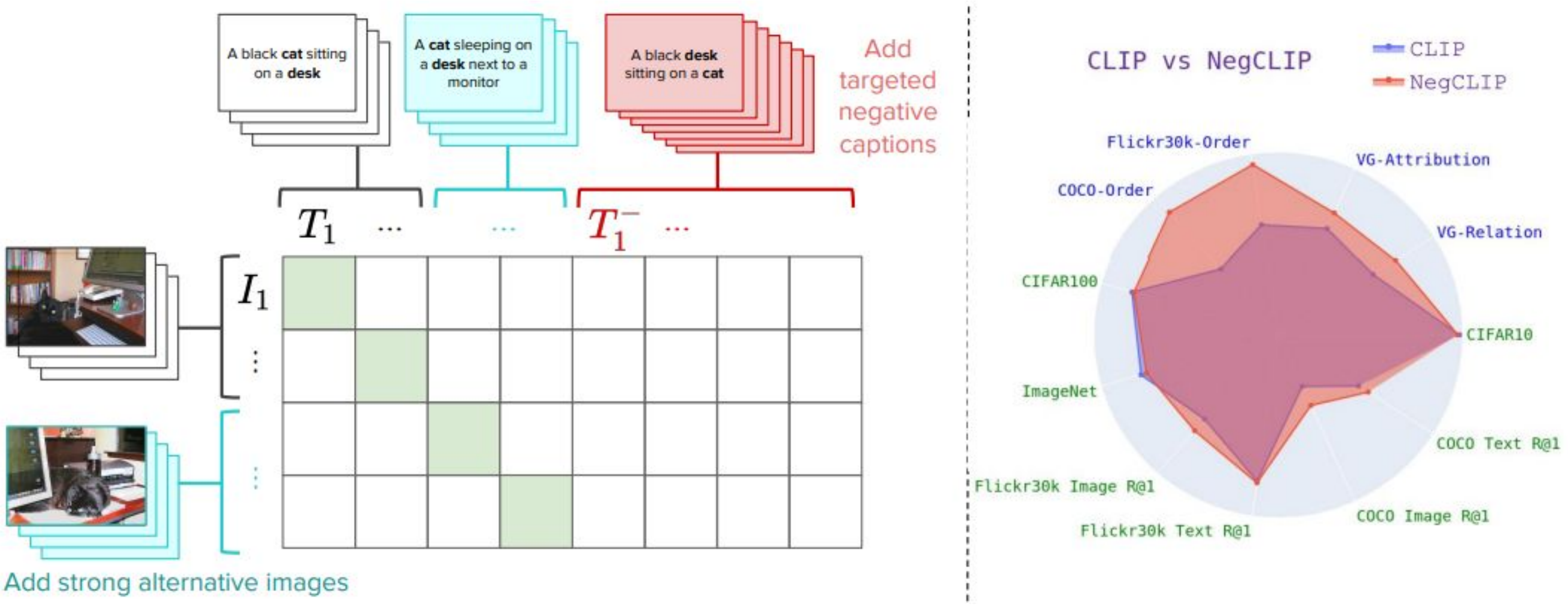
Where to go from here?

- (1/2) Should we work directly in VLMs' conceptual spaces?



Where to go from here?

- (2/2) Should we work directly in VLMs' conceptual spaces? eg. NegCLIP



Thank you for your $\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$!

Comments?

Diego Calanzone, Filippo Merlo
University of Trento

Presenter: **Diego Calanzone, Filippo Merlo**

Seminar: **Grounded Language Processing**

Academic year: **2023/2024**