Getting aligned on Representational alignment

Pavel Popov

A bit of a story

What is representation and where it can occur

- Broadly speaking, information processing systems create representations in which they describe the world around
    - A clown nose is *round* and *soft*, a bowling ball is *round* and *hard*
    - *Hard*, *soft* and *round* can be a few of a many concepts of which our representation of the world consists of
    - Representations are not universal and can vary from system to system
        - the ideas of *sky* is different for birds and humans

What is representation and where it can occur

- Different systems are capable of forming different representations
  - Humans create different semantic neighbourhoods in different languages
  - Humans and monkeys can have homological brain connectivity in response to the same task
  - Teacher and student machine learning models can form similar representations despite different complexity

What can we do with representations

We may be interested in:

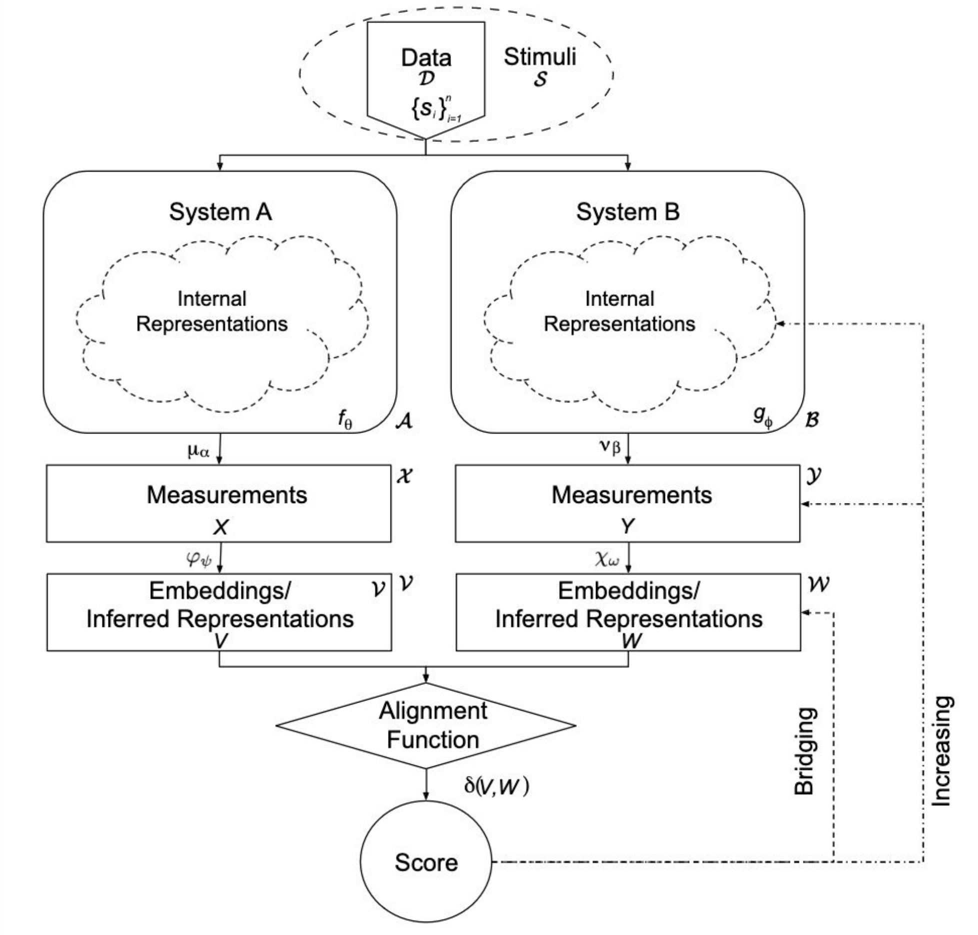
- Measuring the representation similarity of two systems
  - e.g., comparing musical priors across people from different cultures,
- Bridging the representations
  - e.g., projecting the embeddings of visual and language models into a joint space
- Increasing the representational alignment
  - e.g., training a student model to behave like a teacher model

What this paper presents

- A general framework for working with representational alignment problems regardless of domain, be it cognitive science, neuroscience, or machine learning.
- A number of use cases based on the previously published works from different fields that show the versatility of the framework.
- A few remaining challenges of representational alignment
    - This is a work in progress, the authors are calling for feedback for the future revisions

The framework

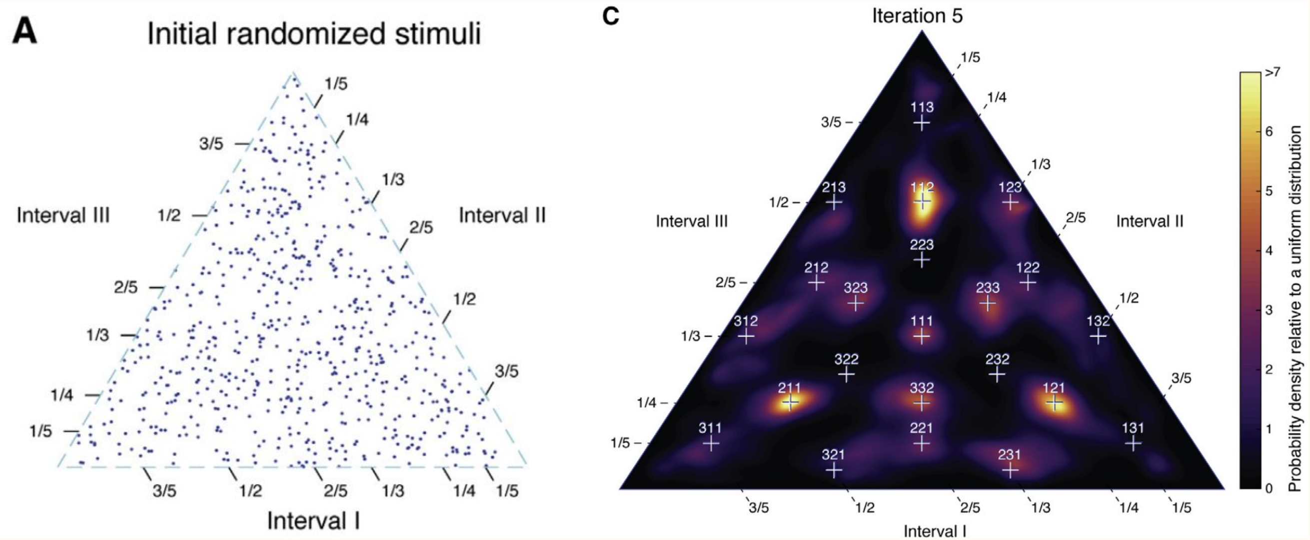
It is a little too technical and abstract.



Examples. Measuring the alignment

Cognitive science: [Jacoby and McDermott, 2017]

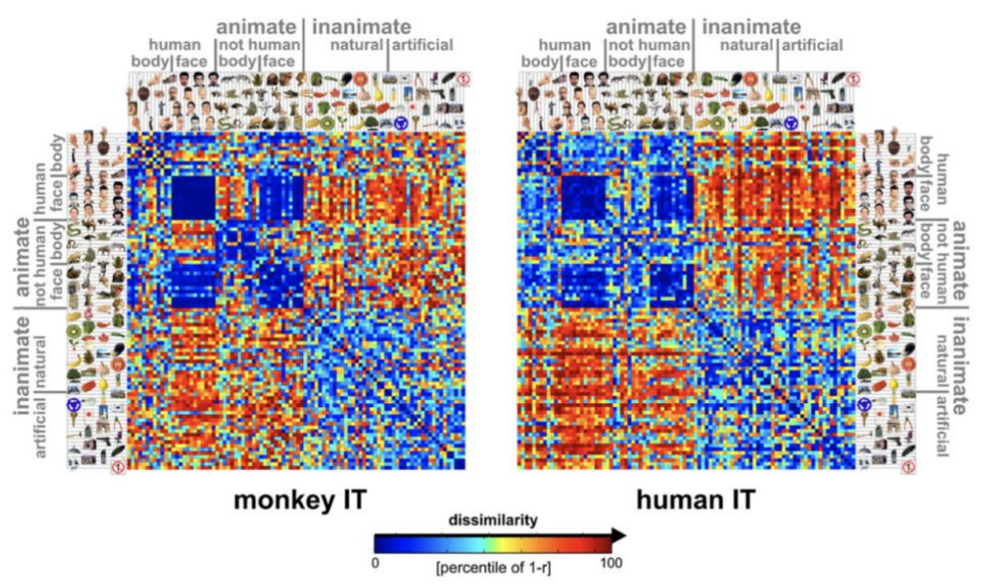
2 groups of people were asked to reproduce a melody made of 3 tones of different lengths. Reproductions were refined iteratively and showed a kind of gravitation map.



Examples. Measuring the alignment

Neuroscience: [Kriegeskorte et al., 2008b]

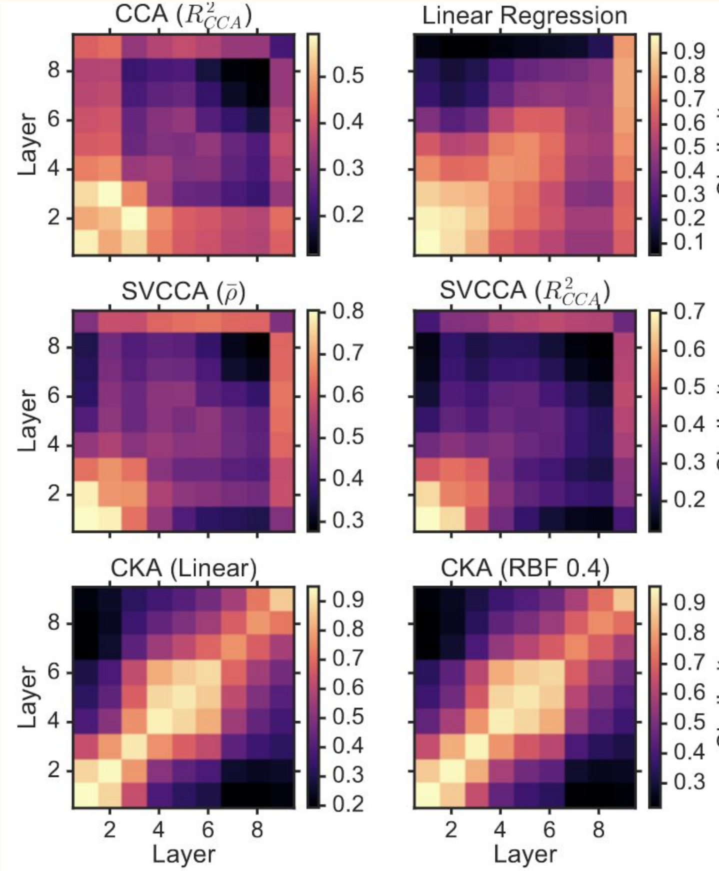
This work measures the alignment between neural responses in monkey and human inferotemporal (IT) cortex.



Examples. Measuring the alignment

Machine learning: [Kornblith et al., 2019]

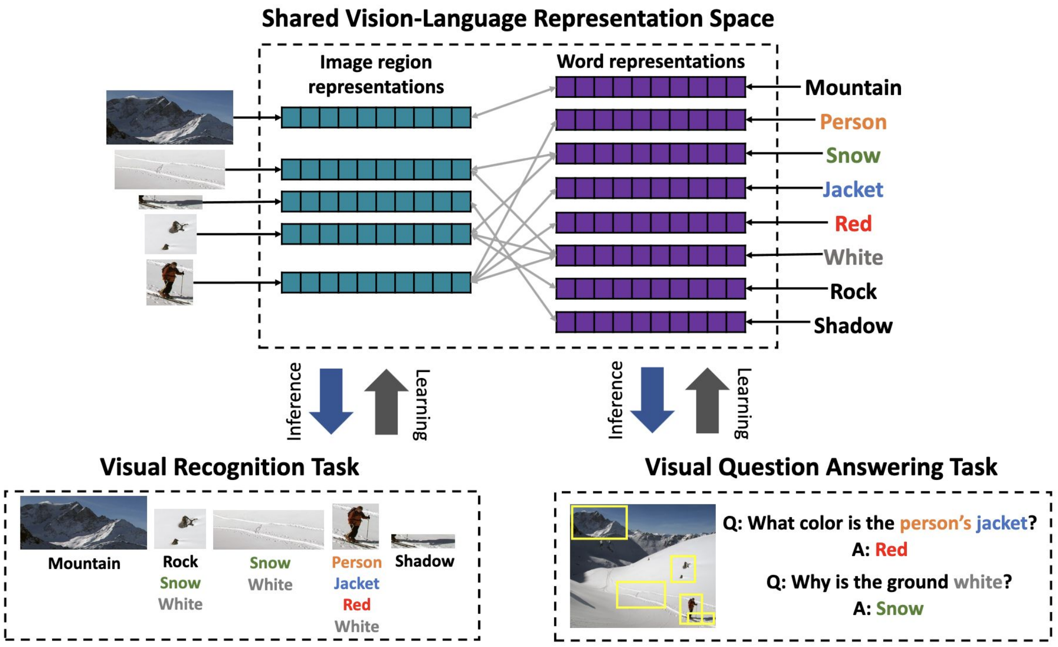
This work tests different similarity measures when comparing activations of different layers of CNN.



Examples. Bridging the representations

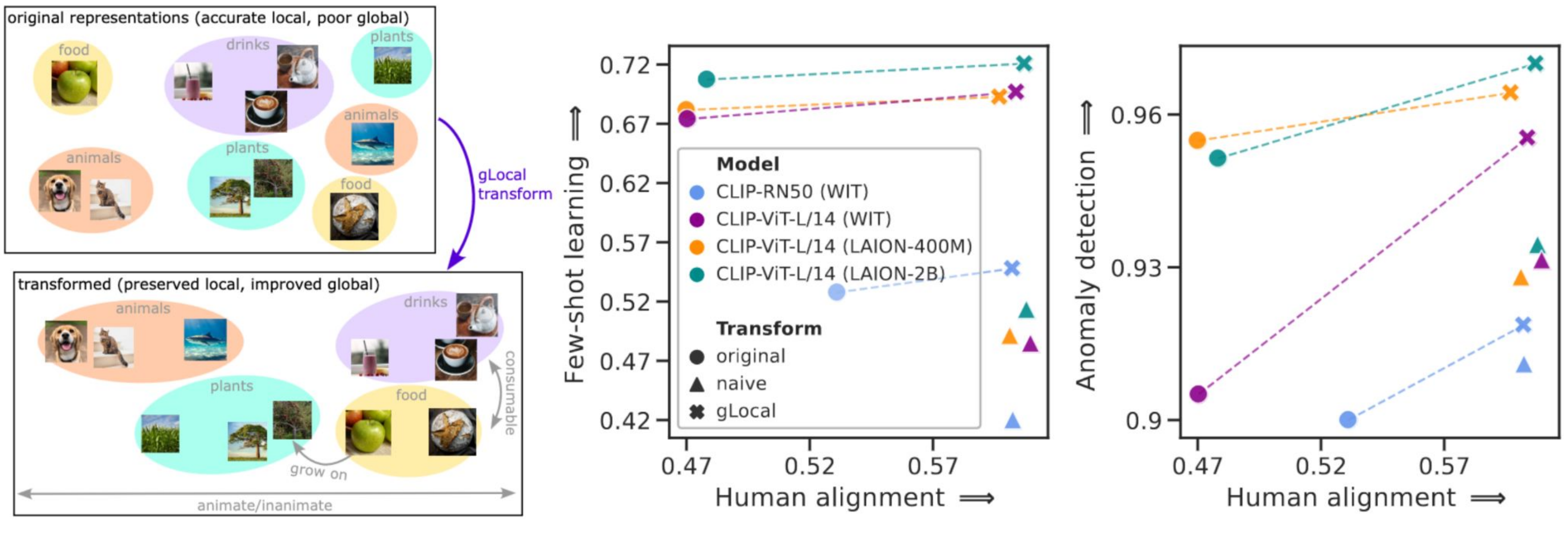
Machine learning: [Gupta et al., 2017]

This work presents a shared vision-language representation space module, which facilitates the information flow between visual and language modules.



Examples. Increasing the alignment

Cognitive science: [Muttenthaler et al., 2023b]. People basically explored an idea how to train a neural net to do the right odd-one-out decisions AND model the decision weights of humans



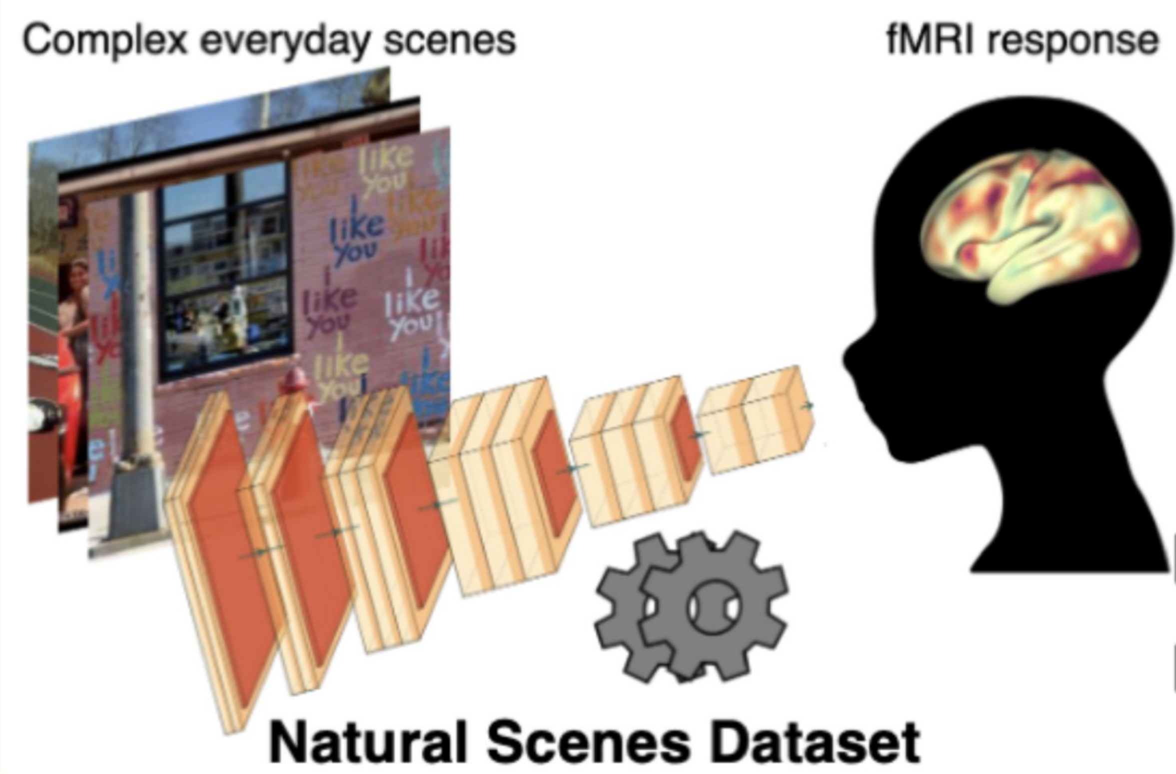
(a) Conceptual cartoon.

(b) Downstream task performance vs. human alignment.

Examples. Increasing the alignment

Neuroscience: [Khosla and Wehbe, 2022]

People trained neural net to model fMRI responses from scenes (not the other way around)



Examples. Increasing the alignment

Machine learning:
Knowledge distillation is one prominent example. It can be used to train simpler student models from complex teacher models, or for multi-modal transfer learning.

For the latter case, [Tian et al., 2019] proposed an objective function, too technical stuff.

$$
\begin{aligned}
\delta(V, W) &= \max_h \mathcal{L}_{\text{critic}}(g_\phi, h) \\
&= \mathbb{E}_{P(X,Y)}\left[\log h(\boldsymbol{x}, \boldsymbol{y})\right] + N\mathbb{E}_{P(X)P(Y)}\left[\log(1 - h(\boldsymbol{x}, \boldsymbol{y}))\right].
\end{aligned}
$$

Background and more

Cognitive science

- Similarity judgments
- Human-machine alignment
- Semantic representations
- Alignment across cultures

Neuroscience

- Individual brains' alignment
- Brain recording
- Brain-model alignment
- Alignment for data enhancement

Machine-learning

- Model-model alignment
- Learning human-like representations
- Interpretability
- Behavioral alignment