BLACKBOX MEETS BLACKBOX

Representational Similarity & Stability Analysis of Neural Language Models and Brains

Samira Abnar, Lisa Beinborn, Rochelle Choenni, Willem Zuidema

If connectionism is to be an adequate theory of mind, we must have a theory of representation for neural networks that allows for individual differences in weighting and architecture while preserving sameness, or at least similarity, of content. In this paper we propose a procedure for measuring sameness of content of neural representations. We argue that the correct way to compare neural representations is through **analysis of the distances between neural activations**, and we present a method for doing so. We then use the technique to demonstrate empirically that **different artificial neural networks** trained by back-propagation on the **same categorisation task**, even with **different representational encodings of the input patterns** and **different numbers of hidden units**, reach states in which **representations at the hidden units are similar**. We discuss how this work provides a rebuttal to Fodor and Lepore's critique of Paul Churchland's state space semantics.

-AARRE LAAKSO & GARRISON COTTRELL (2000)

http://cseweb.ucsd.edu/~gary/pubs/Laakso&Cottrell2000.pdf

"... In order to bridge these divides, we suggest abstracting from the activity patterns themselves and computing representational dissimilarity matrices (RDMs), which characterise the information carried by a given representation in a brain or model. Building on a rich psychological and mathematical literature on similarity analysis, we propose a new experimental and data-analytical framework called representational similarity analysis (RSA), in which multi-channel measures of neural activity are quantitatively related to each other and to computational theory and behaviour by comparing RDMs. We demonstrate RSA by relating representations of visual objects as measured with fMRI in early visual cortex and the fusiform face area to computational models spanning a wide range of complexities... "

–Nikolaus Kriegeskorte,Marieke Mur, and Peter Bandettini (2008)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605405/

Drawing Connections between Computational Models and Biological Models

Given a continues language input, i.e. text of a story:

- How different, are the representations from different models and what can we learn from that…

- How similar are the representations obtained from the models and the activity patterns in human brain

Aligning representational spaces

Learn a Transformation function between the two spaces

- We need enough training data
- We need a proper distance metric between for the target representational space
- What does it mean when we can not learn the proper Transformation function?

Compute Representational similarity

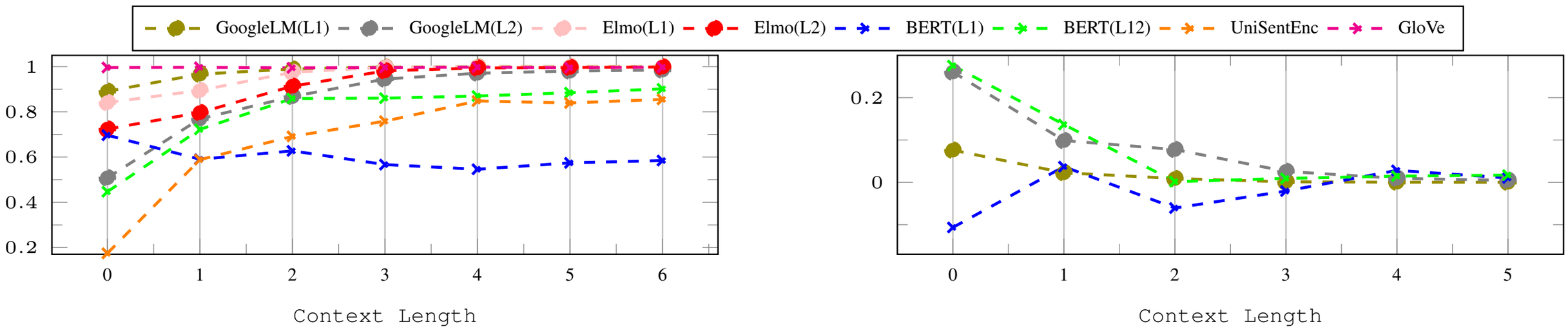
- Similarity of similarities/Distances of distances
- We are interested in the organisation of the entities/concepts in the space rather than the actual activation values of the units
- We assume that we have a fair similarity/distance metric for each of the representational spaces

Representational Stability Analysis

How sensitive the representations are to a specific change in the input condition, e.g. Context Length:

…***witch in the family*** …
…*But for my mother and father, oh no, it was Lily this and Lily that, they were proud of having a* …
…*I was the only one who saw her for what she was -- a freak* …
…
…*How could you not be, my dratted sister being what she was* …
…*"Knew! Of course we knew!* …

$$ReSta(model, k) = RSS(model_{|c|=k}, model_{|c|=k+1})$$



(a) Context Sensitivity ($RSA(L_k\text{-}c_i, L_k\text{-}c_{i+1})$)

(b) Changes in Context Sensitivity ($\delta RSA(L_k\text{-}c_i, L_k\text{-}c_{i+1})$)

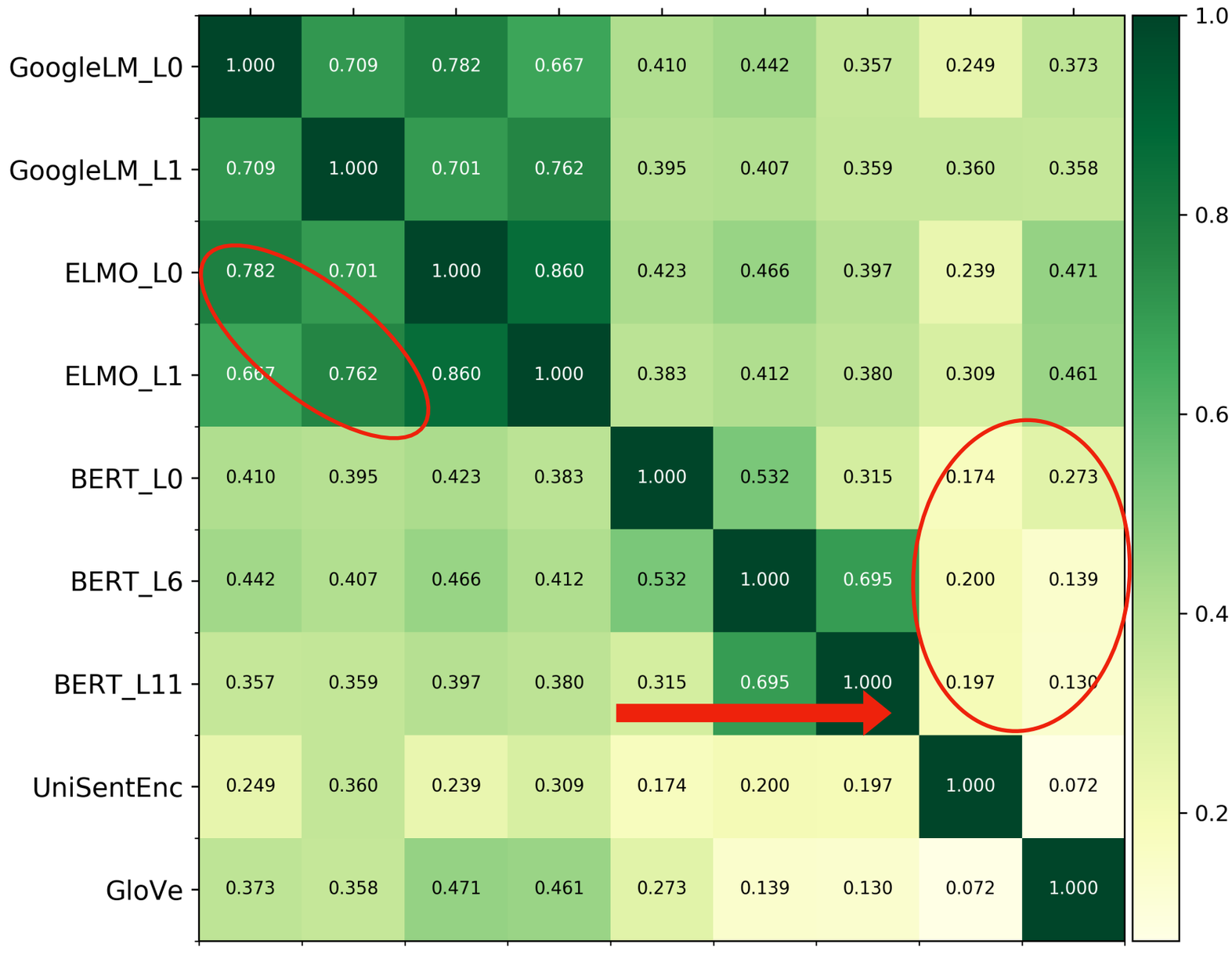
Representational Stability Analysis

(a) GoogleLM

(b) ELMO

(c) BERT

Representational Similarity between Different Models



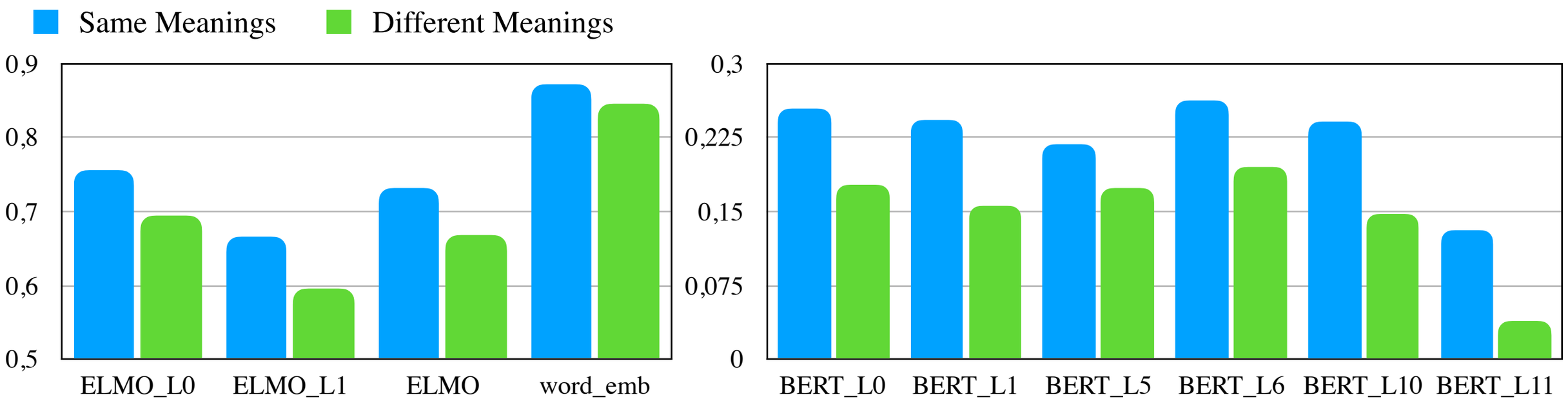
Representational Stability Analysis (evaluation on WiC)

How sensitive the representations are to a specific change in the input condition, e.g. Context:

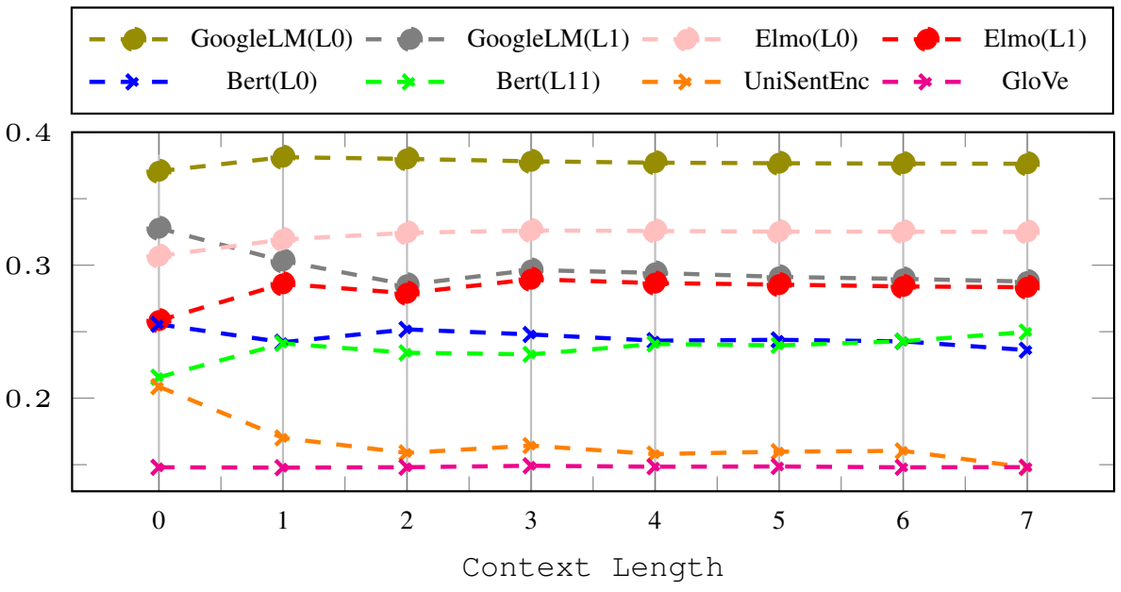
```
He is about average in height .
The snowfall this month is below average .

In the old days a policeman walked a beat and knew all his people by
name .
A beat of the heart .
```

$$ReSt(model, (C_1, C_2)) = RSS(model_{c=C_1}, model_{c=C_2})$$



Representational Similarity between Models and Brains



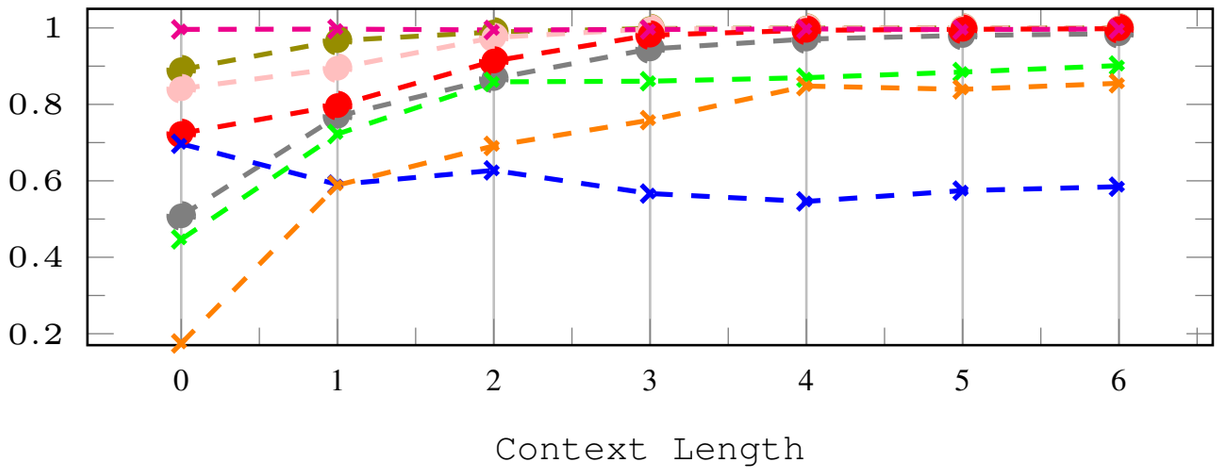
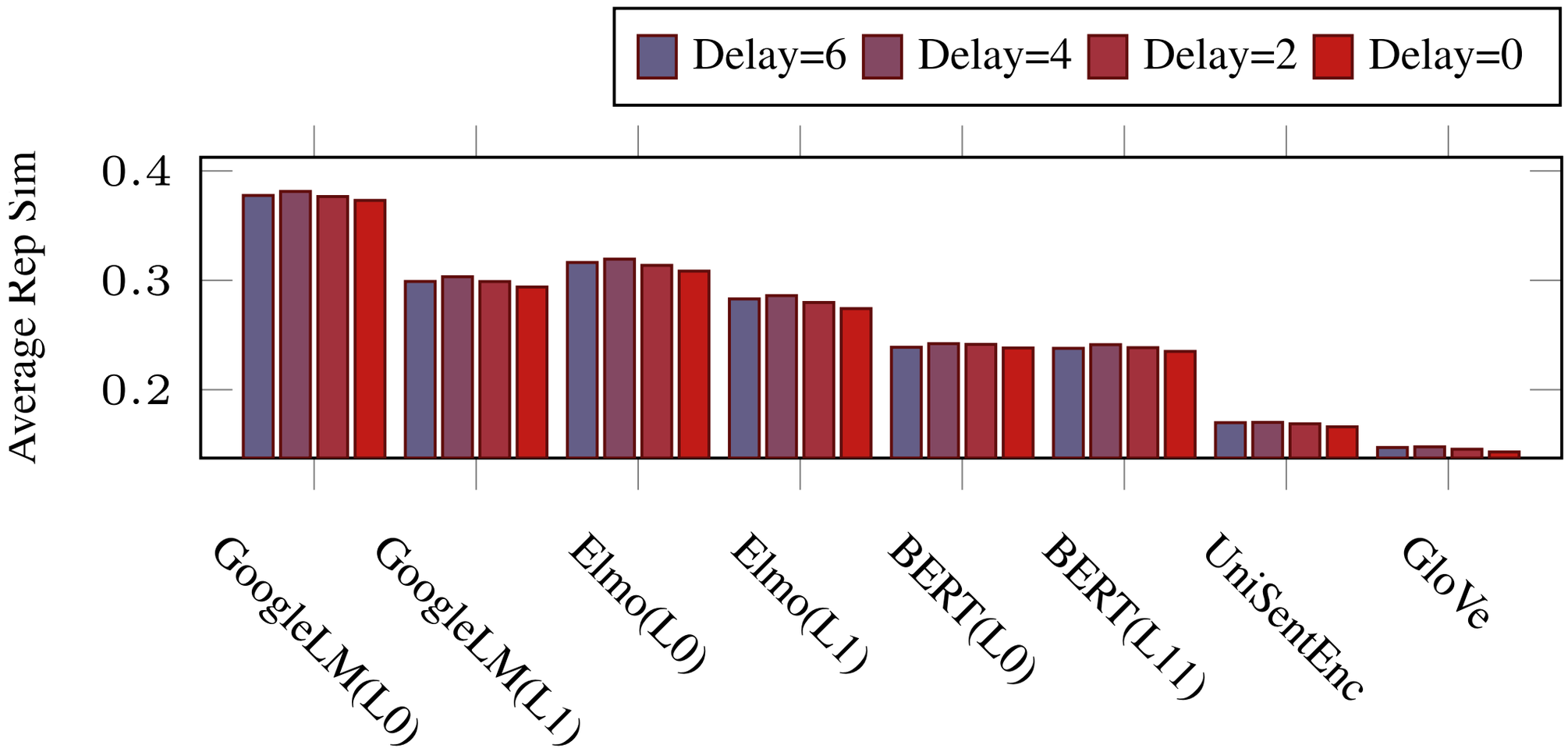


Figure 7: Similarity of the representations from different layers of different models, given different amount of context with brain representations, averaged over all subjects. Note that the average RSS of brains of different human subjects is about 0.55

(a) Context Sensitivity ($RSA(L_k\_c_i, L_k\_c_{i+1})$)

Representational Similarity between Models and Brains



Representational Similarity between Models and Brains