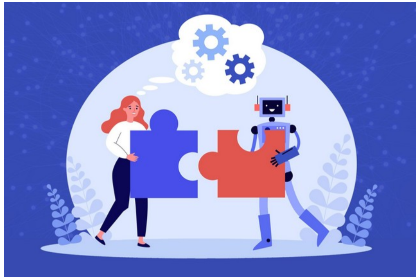


Human alignment of neural network representations

3-Weeks Group Project, 10-28 July, 2023

Jaxartosaurus_Afro - Computer Vision Team 1

Sean Bradley, Fylaktis Fylaktou, Marshall Green, Jarrod Hicks,
George Igwegbe, **Hikaru Tsujimura**

Introduction

Motivations

- Using a similarity judgement task (Figure 1), previous work (Hebart et al., 2020; Muttenhaler et al., 2023) attempted to mimic human-like internal representational space of natural objects on the deep neural nets.

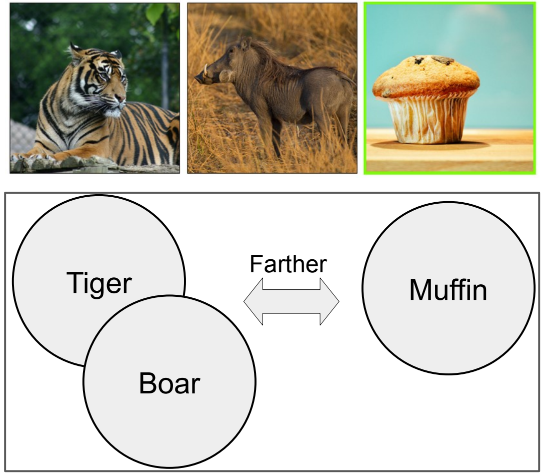
- With large-scale object image (N=1854) and human judgement (N=4.6 millions) datasets and various Deep Neural Net Models (e.g. VGG16, Alexnet) and linear transformation methods, they improved alignment of internal neural representations with human judgments.

- Thus, we planned to replicate with linear transformation methods, and then we used non-linear transformation methods to inspect if there was any difference.

Figure 1

<u>Task</u>: Which is the **least** similar item among the three?

Human choice



Arbitrary human mental space

Hebart et al., 2020


nature human behaviour

ARTICLES

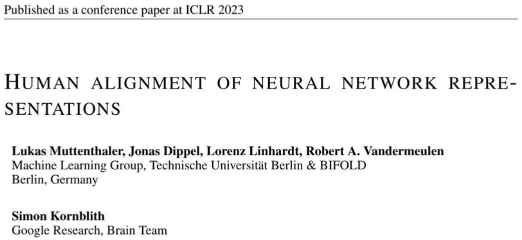
https://doi.org/10.1038/s41562-020-00951-3

Check for updates

Revealing the multidimensional mental representations of natural objects underlying human similarity judgements

Martin N. Hebart[1,2], Charles Y. Zheng[3], Francisco Pereira[3] and Chris I. Baker[1]

Muttenhaler et al., 2023


Published as a conference paper at ICLR 2023

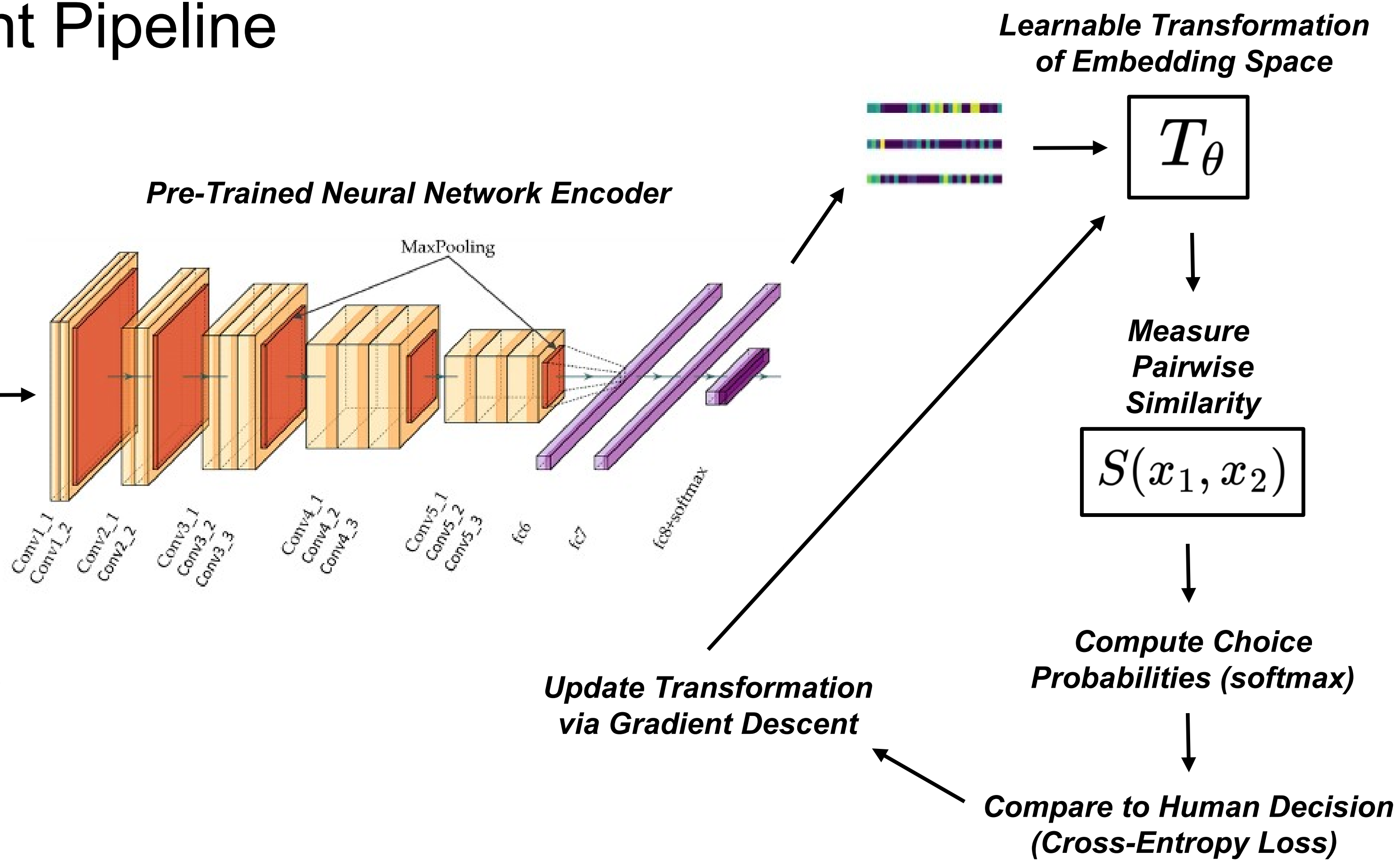
HUMAN ALIGNMENT OF NEURAL NETWORK REPRESENTATIONS

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen
Machine Learning Group, Technische Universität Berlin & BIFOLD
Berlin, Germany

Simon Kornblith
Google Research, Brain Team

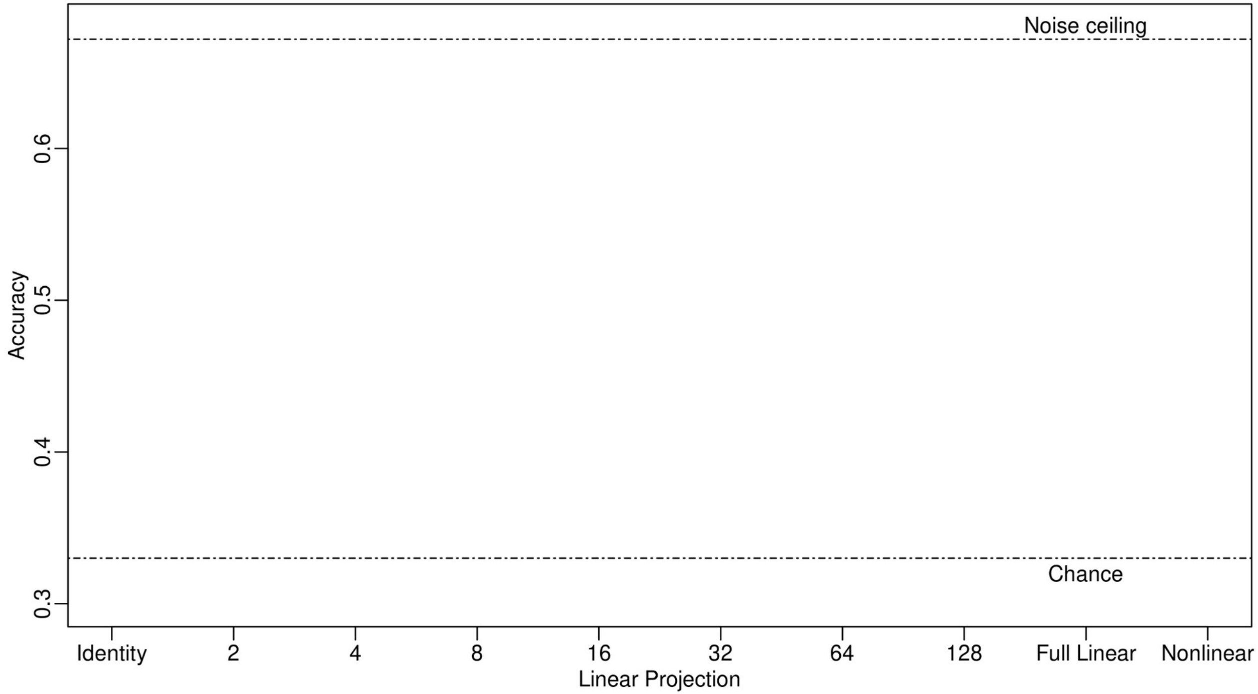
Alignment Pipeline

Figure 2



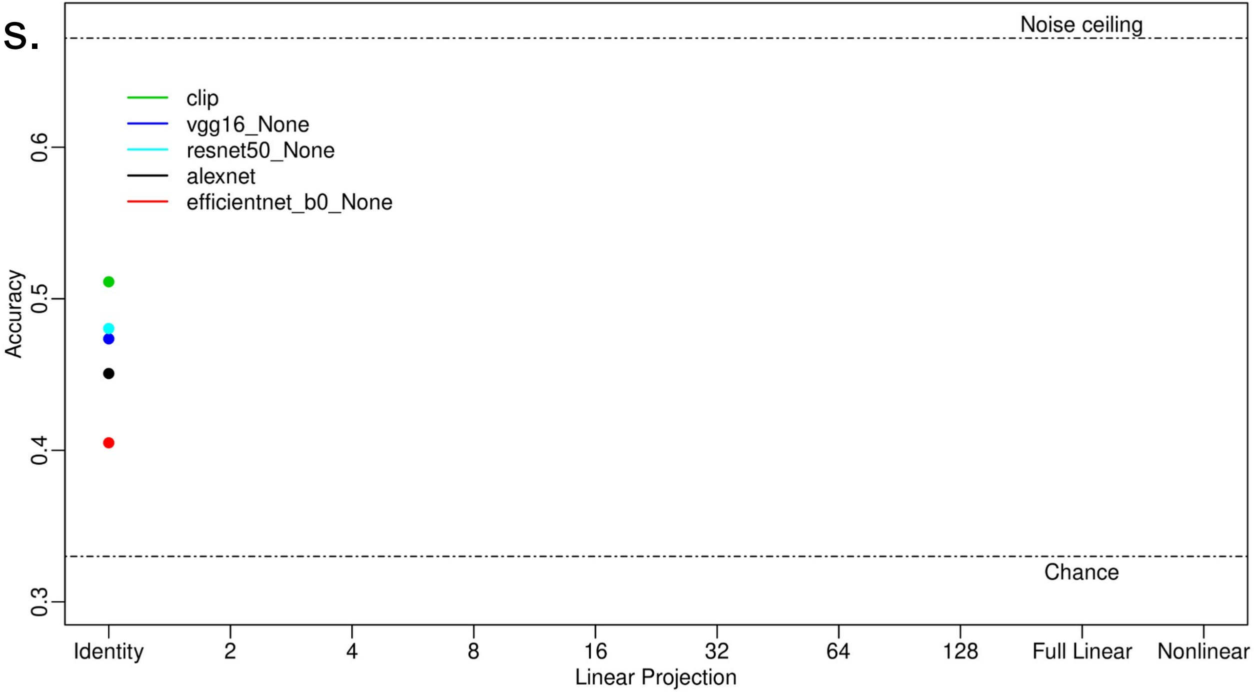
Results: How do the models perform out of the box?

Some perform better than others.



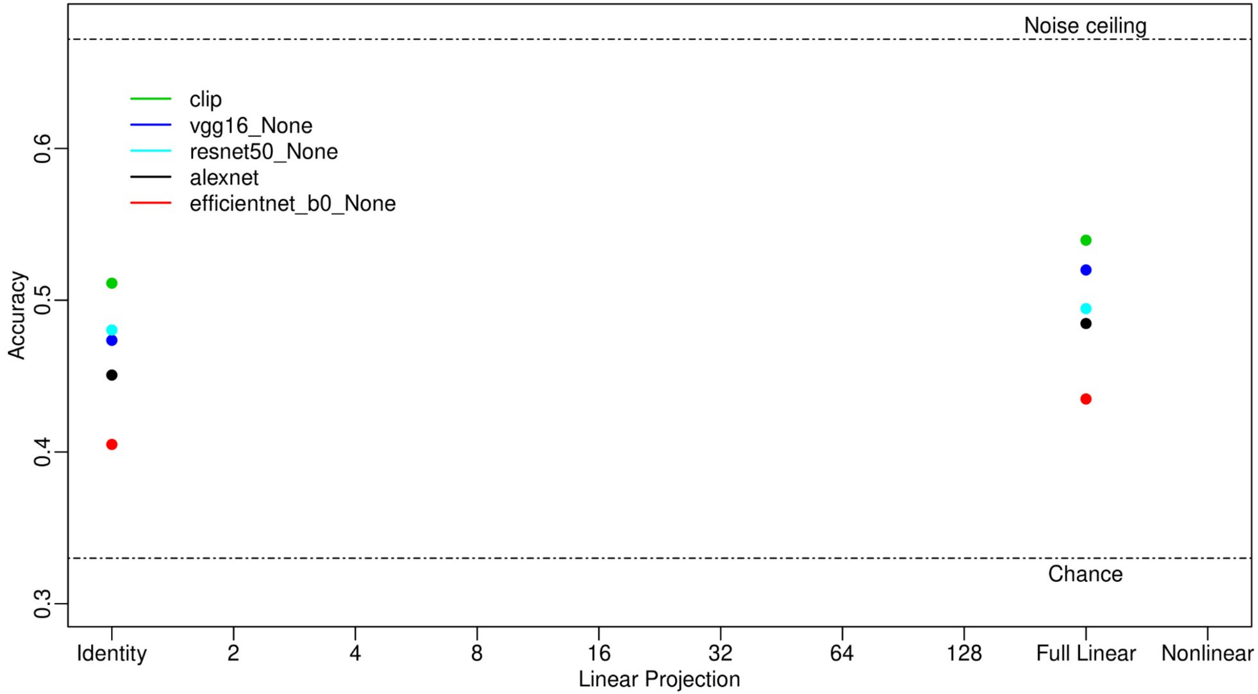
Results: Replication of previous findings

Linear transformation of image representations improves alignment with humans.



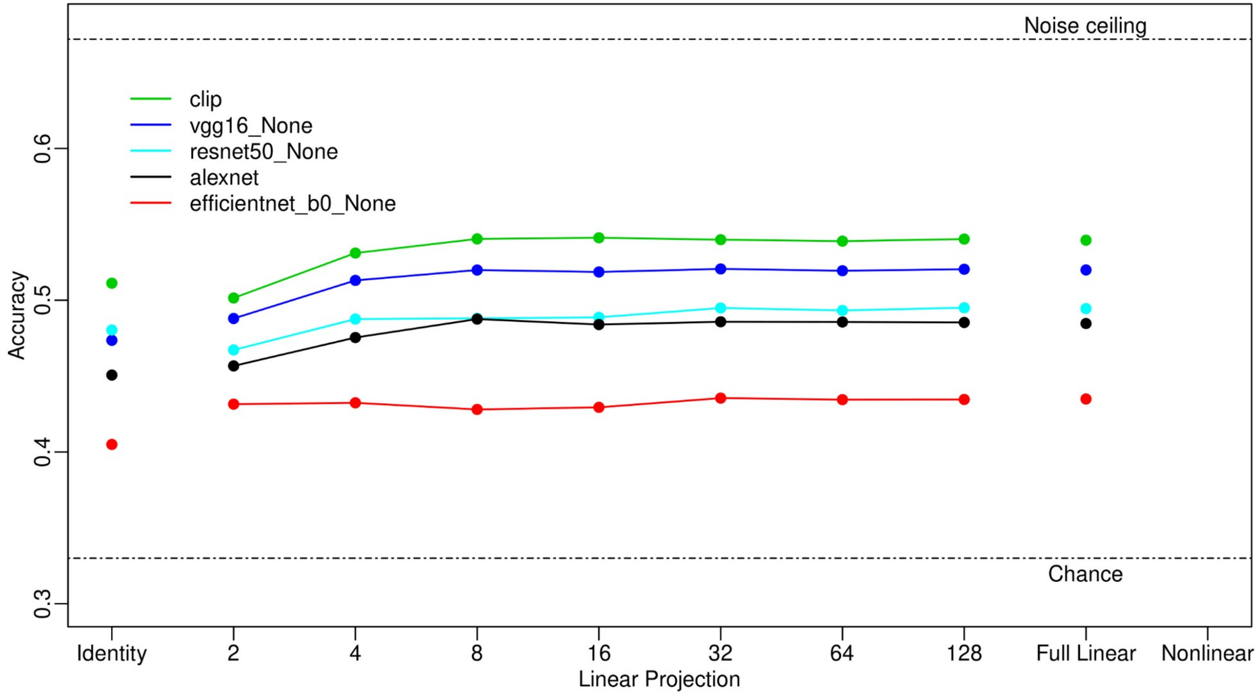
Results: Can we get away with a less complex linear transformation?

Yes! Linear projection with 8 parameters matches full linear transformation



Results: Does a nonlinear transformation improve performance?

Yes! Nonlinear transformation improves all models



Results: Does a nonlinear transformation improve performance?

Yes! Nonlinear transformation improves all models

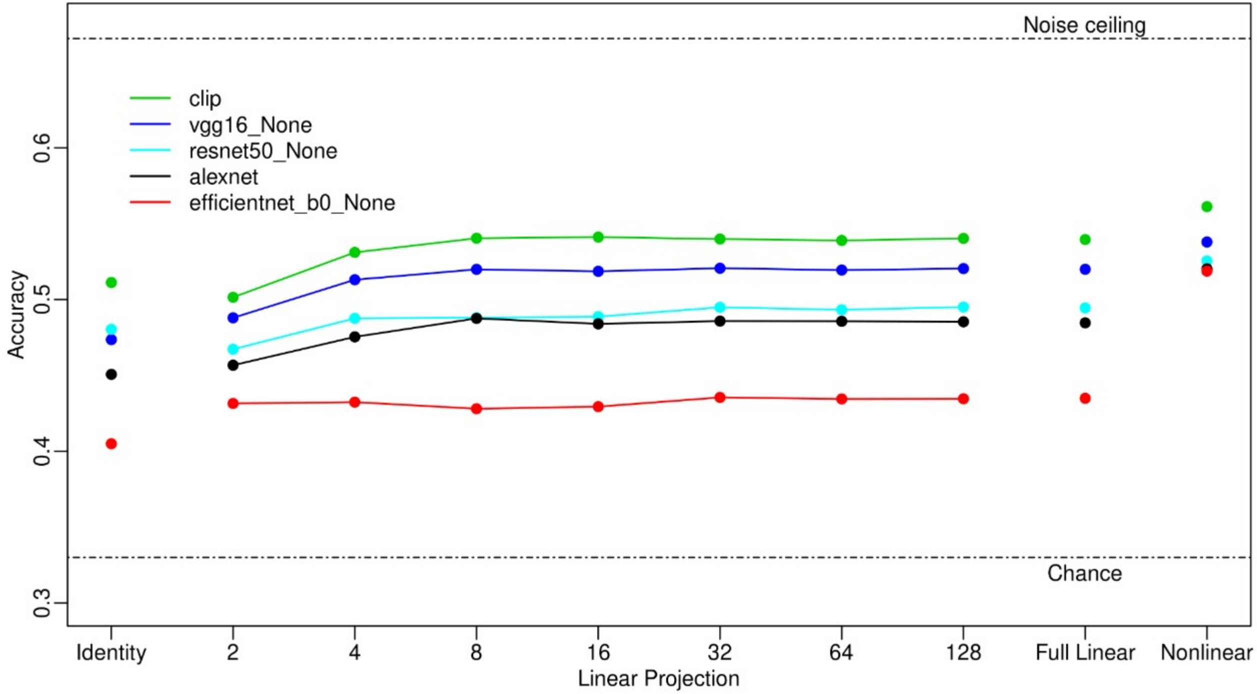


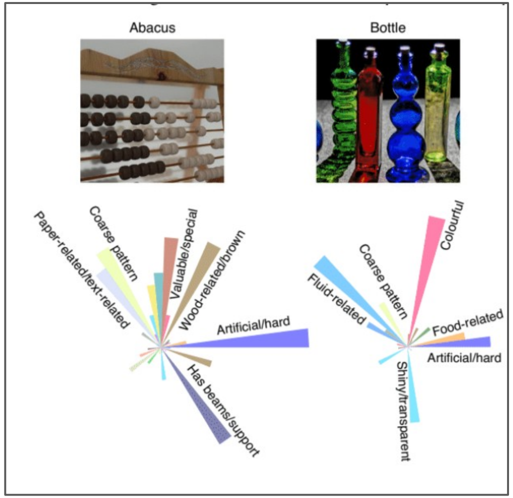
Figure 3

Conclusion

- First, we replicated the original work to show that a linear transformation of artificially-formed neural representations significantly improves alignment with mental representations of humans.
- Then, we examined whether similar non-linear transformations might further improve the alignment with humans.
- Take home message :
    - Both linear- and non-linear transformations of image representation when fed triplets of images improved alignment with human representation in the odd-one-out task, with the majority of the representation being done in the lower dimensions.

Discussion

- A slight majority of zero-shot similarity judgments agree with human raters (~50%)...
- …which isn’t *too bad* given disagreement between humans (~67%!)...
- Most of the model alignment can be represented in surprisingly few features.
- Non-linear transformations achieve better alignment, but there’s tremendous room for improvement.
- What training objectives yield representations that best align with *subjective* human judgments?
- What types of comparisons are easier or more difficult for computers (and people)?



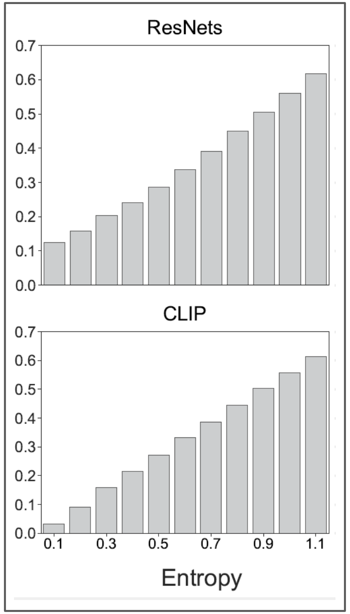
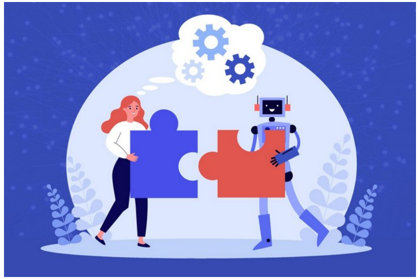


Figure 4



Human alignment of neural network representations - Follow-ups

3-Weeks Solo Project, 31 July – 18 August, 2023

Hikaru Tsujimura

A critical issue – Individual differences

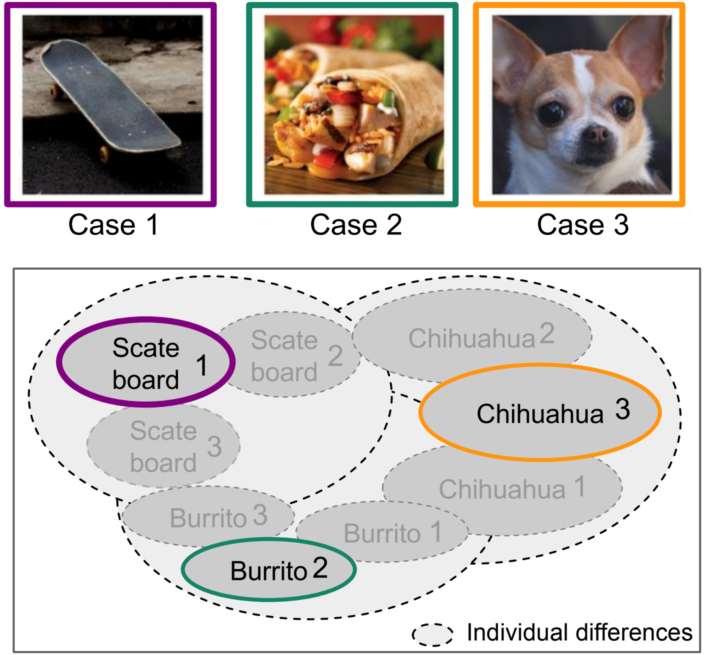
Challenges

With people from different cultures and backgrounds, human judgements vary from person to person (Figure 5).

This wide variation in human judgements makes it challenging to create a single model that accurately predicts human judgements of large-scale populations (Figure 6, red rectangles, less than 60-65% accuracy).

Figure 5

Task: Which is the **least** similar item among the three?

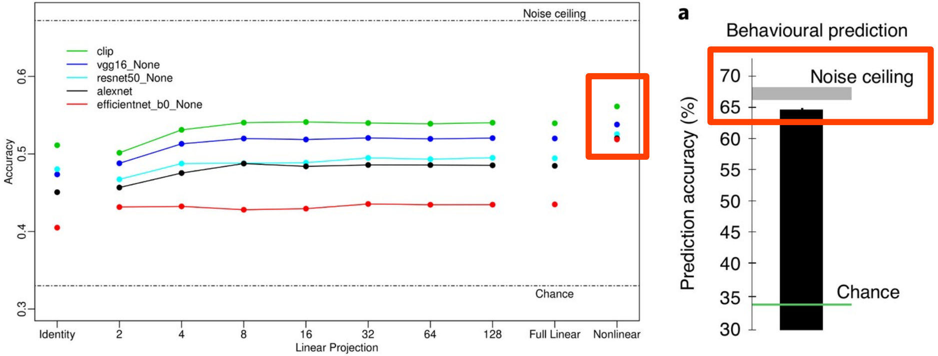


Arbitrary human mental space

Figure 6

Group project

Hebart et al., 2020



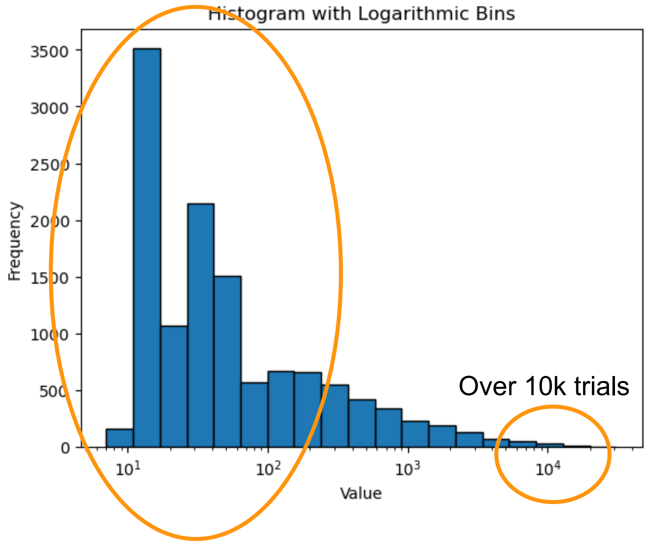
Picking a few subjects with large trials

Attemps

I shifted my research focus towards analyzing data from individuals who completed large trials in this study (top 10 subjects with 15k~ trials), although there are many individuals with ~100 trials.

Instead of aiming to enhance accuracy of a single model across all subjects, I became curious if I can create a specialised model that can accurately predict the similarity judgement of those individuals with a large number of trials better than the previous work.

As you can see the number of trials conducted by each of the top 10 subjects in this task (Figure 7), there is still a big variability in a number of trials. If, even with a lower number of trials (compared to original work with a total of 4.6 millions data), an individually customised model can predict individual's judgement, it's promising that an individually tailored deep neural model is a key to predict human's behaviors and "disagreements" between humans.



Under 100 trials

Figure 7

ID	# of Trials
1	31,033
2	30,683
3	29,884
4	24,125
5	22,843
6	20,870
7	20,285
8	19,047
9	19,033
10	17,791

Methodology

Alignment pipeline

I used similar methods as the group project. Yet, it's a solo project with limited man power and resources, so I focused on fixed parameters, and only using one type of pre-trained network encoder (word2vec vectors based on the google 300 news (https://code.google.com/archive/p/word2vec/)) applying to deep neural models and add an additional neural model of embedding layers to see if semantic (high-level) representations are unique to each individual.

I used the google 300 news vectors and an embedding model because I can visualise each person's semantic representations and unique relationships between words within each individual, similar to results across all subjects of the original work (Hebart et al., 2020) or other previous work with emoji symbols (Eisner, et al., 2016), by using the t-SNE technique.

Check page3

Word2vec, Google news vectors

Alignment Pipeline

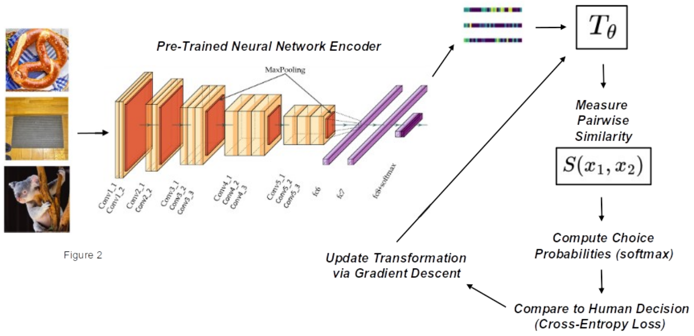
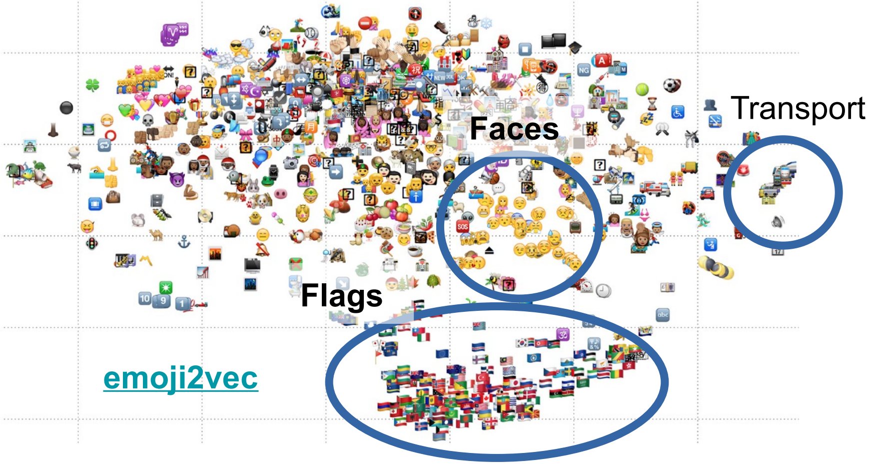


Figure 2



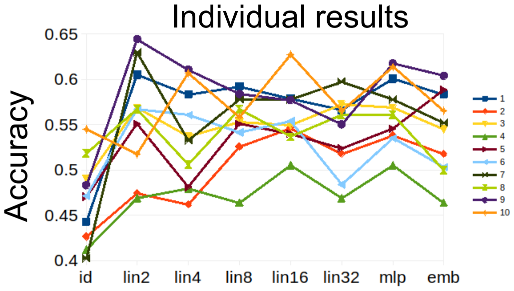
[emoji2vec]()

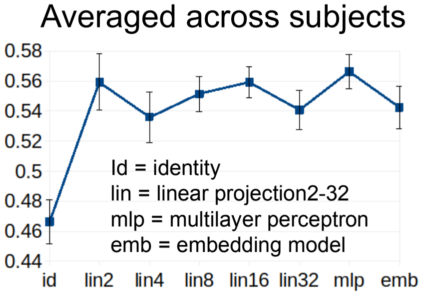
Results - 1

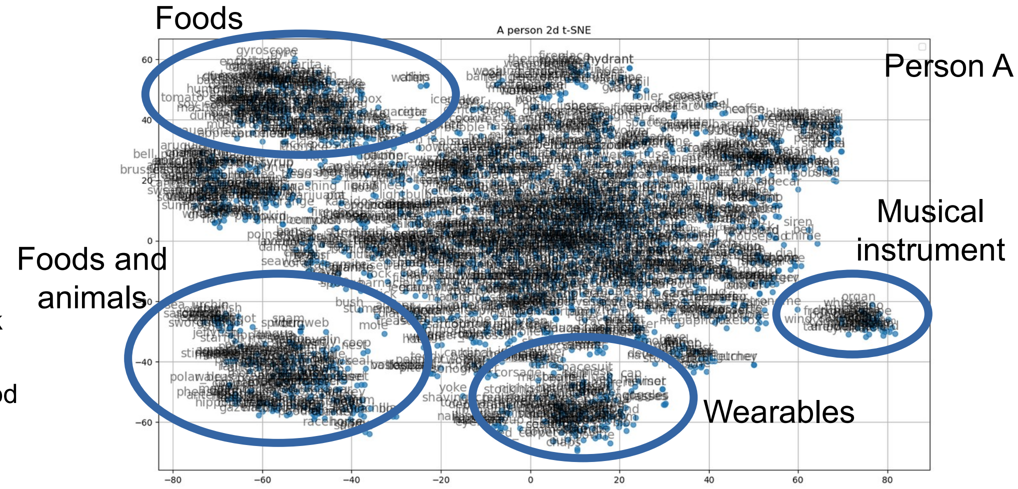
Accuracy and semantic representations

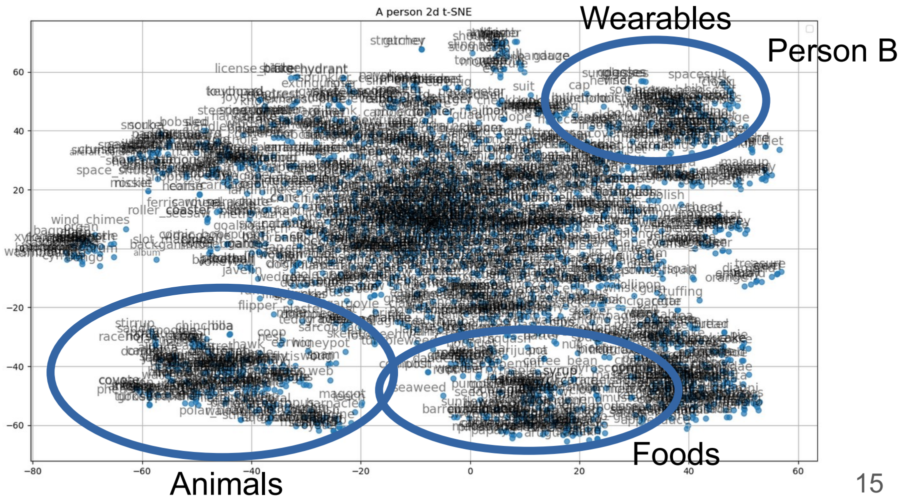
According to initial results (bottom), even based on data of 15k-30k trials (0.3% of 4.6 millions' original data), each individually customised model can accurately predict human judgement as good as original study did. This suggests that a customised model for predicting an individual's behaviors will be critical.

Also, t-SNE technique (right) is useful to visualise semantic representations and relationships between 1652 words of each subject and shows some distinct and similar representation patterns each other.





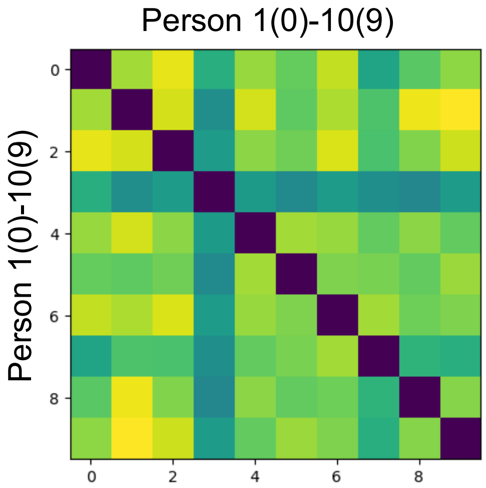




Results - 2

Similarity of semantic representations between subjects is poor

Yet, with the semantic representations of each subject with 1652 words are not useful to detect whose semantic represenation is similiar to whom. That is probably because the word vector dimension is too large for each participant (i.e. 1652 words x 300 google news vector = around 500k vectors per subject). This indicates one of the major reasons why it is not easy to create a single model to represent an universal semantic representation to predict human's judgement in a large population. Per individual words, representation is variable.



Cosine similarity distance
- median 0.03143944963812828
- max 0.040745481848716736
- min 0.0

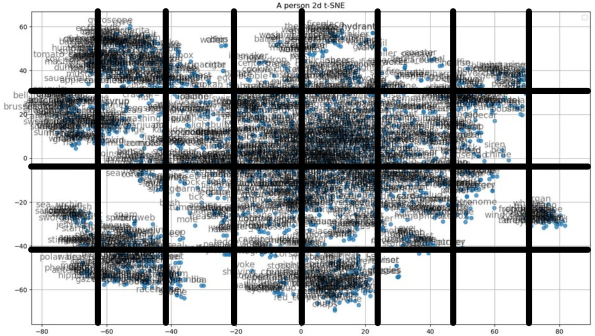
Results - 3

Similarity of higher-level semantic representations between subjects is better

Therefore, I preprocssed the 2D mental representations of 1652 individual words into higher semantic representations for each subject, by using a technique of Kernel Representation (Butnaru and Ionescu 2018). That is, At first, I divided 1652 words in the t-SNE 2D map into a 16x16 dimensions, and then aggregated into 8x8 dimension, which turns into 64 semantic cells with 300 vectors each. As a result, a final vector size of each individual is reduced to 64x300 = around 19k. With the 19k vectors per subject, I achieved to represent each individual's semantic space (like the gray square on the right), and then, that identifies that some semantic representations betwen two people is highly similar (e.g. Person 1 and 8, Person 3 and 6 on the bottom square) while semantic representations between other people pairs are dissimilar (e.g. Person 1 and 9, Person 2 and 4-7).

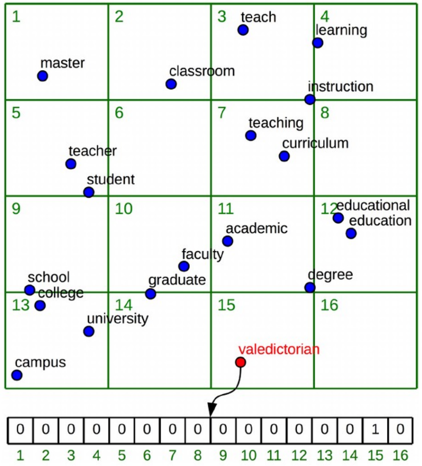
I call the long 1x19k vectors as a "Person2vec".

A person's 2d mental representations

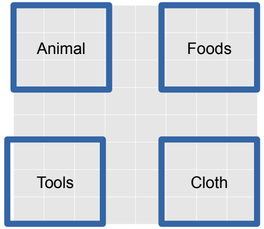


16x16 dimension = 256 cells

Kernel Representations
(Butnaru and Ionescu 2018)

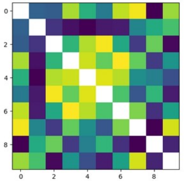


Feature extractions



Vectorizing

8x8 dimension = 64 cells



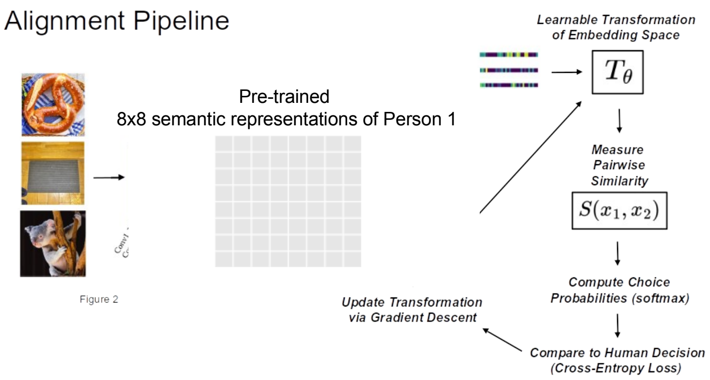
Cosine similarity distance
- median 0.13470104336738586
- max 0.9947574138641357
- min -0.9945352077484131

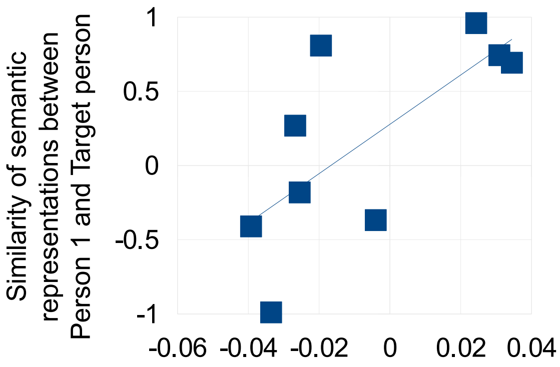
Results - 4

Transferrable embedding space

After identifying similar semantic representations between certain people pairs and dissimilar semantic representations with other pairs, I checked if such representation similarity indicates that transfer learning between similar semantic representations are effective.

Similar to application of pre-trained neural network encoder, semantically similar representation of person A to person B is effective for transfer learning, indicating by improvement of better accuracy with pre-trained representation, while dissimlar semantic representations make it worse.





Accuracy improvement (positive value) and deterioration (negative value) after applying pre-trained semantic representations of Person1 to Target person

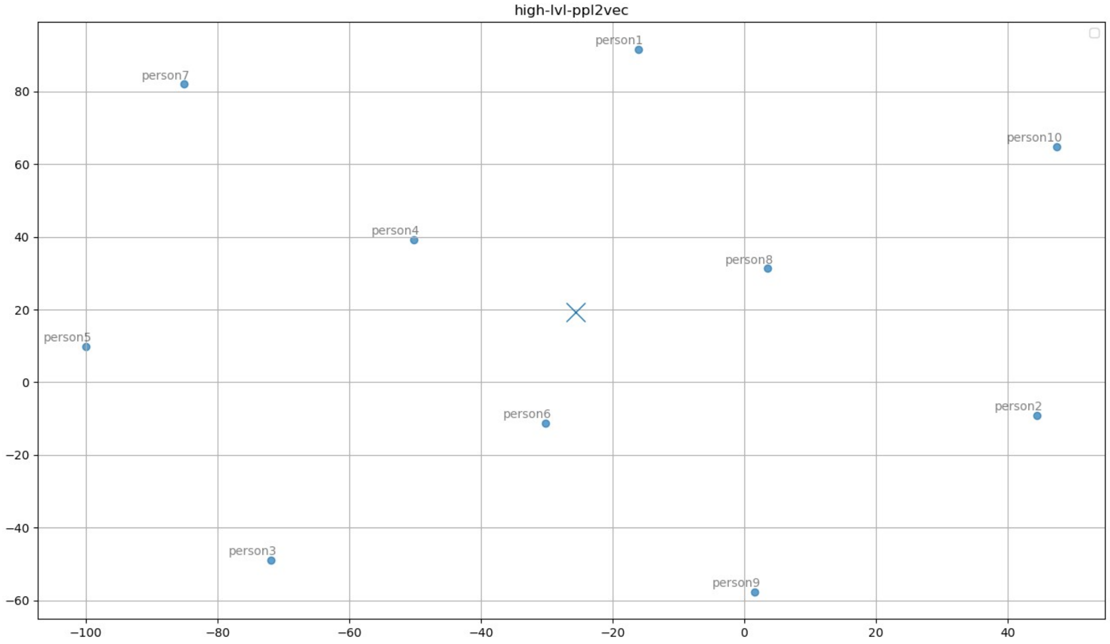
People2vec

Next challenges

By representing each person's high-level semantic vector in a same space, in a similar manner as the t-SNE 2D space with more than 100 or more subjects, I should be able to draw a people2vec for semantic representations. This way, transferable and sharable representations can be aggregated into a cluster of particular people type (e.g. Animal-centred person, food-centred person, action-centred-person, etc). By correctly classifying each person into a certain people cluster, I believe that accuracy will be improved and with larger populations.

What is interesting in the datasets in this particular study is that, how accurate such particular people cluster can accurately predict behaviors of a person with only a few trials avaialble, such as 20-100 trials.

Example people2vec representation



Conclusion, discussion and future directions

Summary

- As a take-home summary, individually-tailored semantic representations are very useful to accurately predict target person's behaviors, while a single model predicting entire human populations might face a disagreement between humans, as well as visualising semantic representations in a 2D map.

- At the same time, individually-customised semantic representations are an effective path to aggregate and share with similar semantic represenations of other people.

- Similar semantic representations are helpful for transfer learning.

- Furthermore, this deep learning method is additive so that adding more data will help to improve accuracy of this model and understand more accurate representations of each individual and generalized semantic representations of a certain group of people (*not all populations), which is deep-learning friendly!

- Future direction will challenge how to implement a model or apply to individuals with a few trials, such as those with ~100 trials and require implementations of more sophiscticated deep neural architectures to perform this task more accurately (e.g. 80-90%).

Supplementary

Minimum # of trials

I checked performance of my model on the 115 subjects with large # of trials completed. Accuracy here means that an overall accuracy for the linear projections to the embedding layer model. Thus, my model works very well on some people (over 70%), but not good enough for some people (under 50% compared to the original and group project).

I would say, my model works up until around top 30 subjects with 10k~ trials. This will be a challange for me to improve a model with better accuracy with lower number of trials in my next goal.

of completed trials

Subject #															
1-15	31033	30683	29884	24125	22843	20870	20285	19047	19033	17791	16752	16083	15260	15189	14500
16-30	13276	12403	12268	12099	11626	11432	11354	11313	11312	11045	10829	10649	10353	10302	10219
31-45	9850	9828	9810	9656	9583	9385	9106	8978	8897	8867	8655	8631	8617	8604	8575
46-60	8572	8452	8319	8239	8177	8135	8072	8005	7811	7716	7461	7434	7413	7321	7222
61-75	7132	7090	7027	6976	6806	6696	6679	6567	6564	6558	6494	6489	6333	6317	6274
76-90	6203	6164	6031	5992	5911	5906	5906	5873	5861	5822	5771	5678	5648	5624	5582
91-105	5555	5531	5498	5455	5444	5416	5398	5395	5394	5379	5328	5317	5309	5264	5257
106-115	5239	5174	5159	5101	5046	5033	4991	4971	4836	4815					

Accuracy

Subject #															
1-15	0.58	0.52	0.55	0.48	0.54	0.53	0.57	0.54	0.59	0.59	0.64	0.57	0.61	0.44	0.7
16-30	0.47	0.5	0.6	0.43	0.68	0.6	0.58	0.52	0.59	0.47	0.58	0.46	0.67	0.67	0.69
31-45	0.52	0.45	0.61	0.72	0.74	0.34	0.67	0.46	0.64	0.45	0.5	0.63	0.47	0.37	0.3
46-60	0.73	0.65	0.66	0.57	0.58	0.53	0.41	0.57	0.49	0.55	0.49	0.37	0.63	0.42	0.42
61-75	0.4	0.36	0.43	0.62	0.55	0.46	0.54	0.52	0.54	0.37	0.43	0.36	0.59	0.46	0.54
76-90	0.61	0.59	0.56	0.56	0.58	0.57	0.61	0.59	0.48	0.59	0.59	0.47	0.7	0.62	0.61
91-105	0.52	0.44	0.42	0.53	0.5	0.41	0.63	0.54	0.52	0.47	0.51	0.57	0.65	0.53	0.4
106-115	0.5	0.52	0.62	0.47	0.6	0.5	0.62	0.44	0.56	0.31					