Reassessing Number-detector Units in Convolutional Neural Networks

Nhut Truong, Shahryar Noei, Alireza Karami

NeurIPS Workshop on Behavioral Machine Learning

December 14, 2024

UNIVERSITÀ DI TRENTO | CiMeC Center for Mind/Brain Sciences

FONDAZIONE BRUNO KESSLER

NeuroSpin

Nhut Truong

PhD student

Center for Mind/Brain Sciences (CIMeC)

University of Trento

leminhnhut.truong@unitn.it   tlmnhut.github.io

Background

- **Numerosity** - the ability to perceive and estimate the number of items in a visual scene - is believed to be represented by **"number-detector" units** within Convolutional Neural Networks (*Nasr et al., 2019; Kim et al., 2021*)

- However, *Karami et al. (2023)*, using Representational Similarity Analysis (RSA) demonstrated that **CNNs fall short of explaining the variance in numerosity representation** observed in the brain.

Background

- The classical RSA framework (*Karami*) assumes equal contribution of **all features**, which can **underestimate** the correspondence between models and behavior data.

- Moreover, this approach may **overemphasize irrelevant features**, potentially **overlooking behaviorally relevant information** like number-detector units.

Our contribution

- We used a **pruning approach** to identify units in CNNs that **best represent numerosity** at the population level and improve **alignment with behavioral** data.

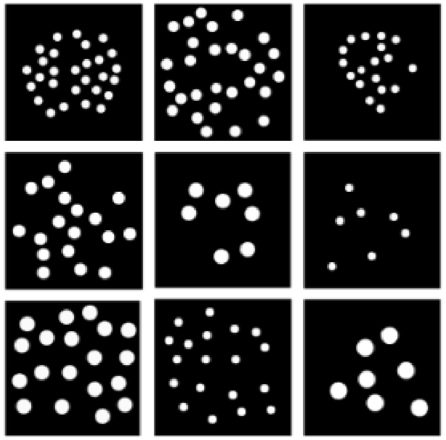
- Pruning **removes the redundancy** in pretrained models, retains only the **most relevant units** for numerosity representation.

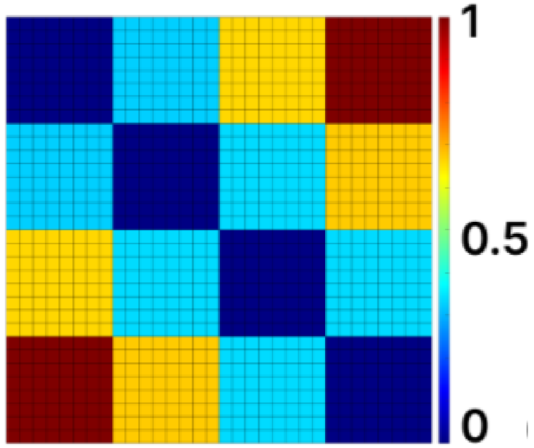
Models and stimuli

- Models:
  - Pretrained CORnet-Z and CORnet-S (*Kubilius et al., 2018*).
  - 3 versions: trained on ImageNet, trained for numerosity discrimination (DeWind et al., 2015), and untrained.
  - Target layers: V1, V2, V4, IT.

- Stimuli: Visual dot sets with varying numerosities and visual features.

- Behavioral number RDM: simulated logarithmic distance between the pairs of condition.





Pruning method

- Pruning (*Tarigopula et., 2023*) involves 3 steps:
  1. **Importance Assessment**: Each unit is individually removed, and the resulting RDM is compared to the number RDM. Significant drops in score indicate important units; smaller drops or increases suggest unimportant or noisy units.
  2. **Ranking**: Units are ranked from most to least important based on their impact on the RDM score.
  3. **Sequential Reintroduction**: Units are reintroduced in ranked order, and the RDM fit is reevaluated after each addition. The process stops when the highest RSA score is achieved, defining the "retained units."

- Compare with:
  - **Full (unpruned) model**
  - **Number-detector units** identified via ANOVA (*Nasr et al., 2019; Kim et al., 2021*)

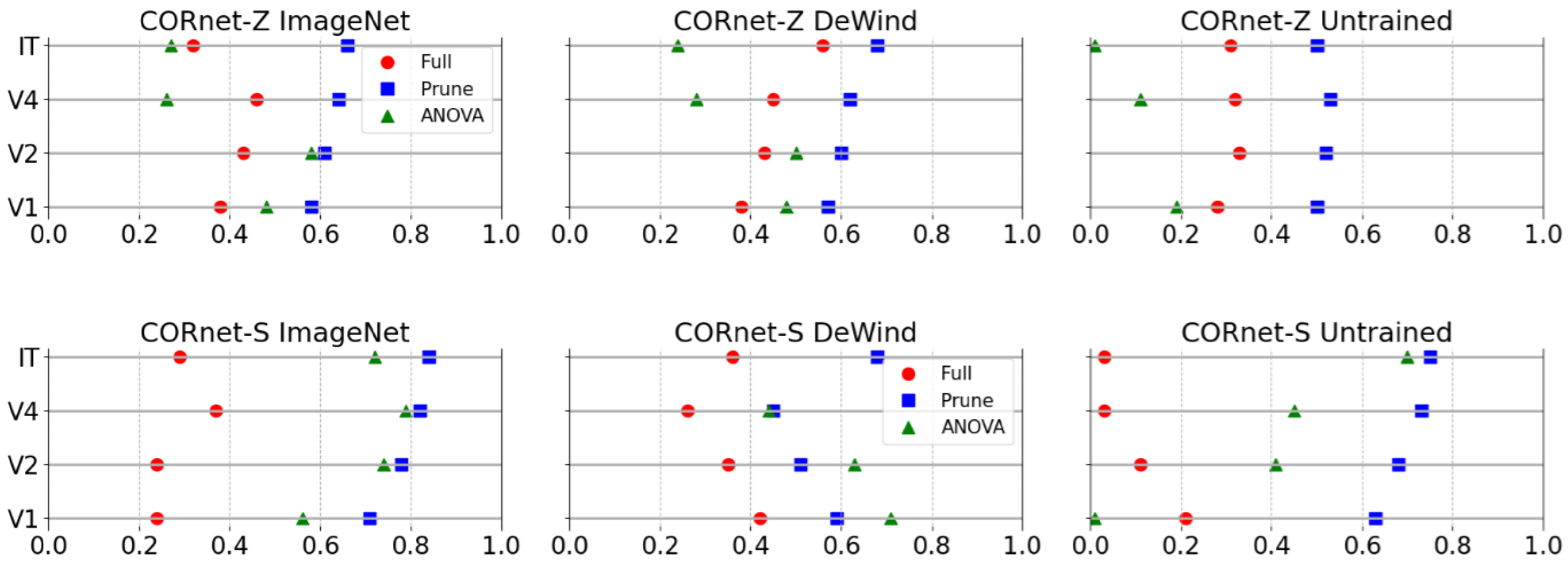
Retained Units and Number-Detector Units Often Do Not Overlap

- Little to no overlap was observed in the IT layer of both models and in the V4 layer of CORnet-S.
- Significant overlap was found in the V2 and V4 layers of CORnet-Z, and in the V1 layer of CORnet-S.
- Only 3 cases showed a perfect overlap score of 1, while 7 cases had a score of 0.

CORnet	Layer	ImageNet	DeWind	Untrained
Z	V1	0.40	0.57	0
	V2	0.59	0.71	1
	V4	1	0.85	1
	IT	0.09	0	0
S	V1	0.31	0.27	0.65
	V2	0.01	0.21	0.01
	V4	0	0	0
	IT	0	-	0

Retained Units Fit the Behavior Data Better than Number-Detector (ANOVA) Units

RSA Pearson correlations



Conclusions

- Using RSA on pruned models, we tested if traditional number-detector units in CNNs can capture numerosity
- **The results show that number-detector units in CNNs are not essential for numerosity representation.**
- Future directions include using explainable AI to decode selected units, exploring more naturalistic datasets, and extending analyses to the language domain.

References

- Nasr, K., Viswanathan, P., & Nieder, A. (2019). *Number detectors spontaneously emerge in a deep neural network designed for visual object recognition.* Science Advances, 5(5). https://doi.org/10.1126/sciadv.aav7903
- Kim, G., Jang, J., Baek, S., Song, M., & Paik, S. -B. (2021). *Visual number sense in untrained deep neural networks.* Science Advances, 7(1). https://doi.org/10.1126/sciadv.abd6127
- Karami, A., Castaldi, E., Eger, E., & Piazza, M. (2023). *Neural codes for visual numerosity independent of other quantities are present both in the dorsal and in the ventral stream of the human brain.* bioRxiv (ColdSpring Harbor Laboratory). https://doi.org/10.1101/2023.12.18.571155
- DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). *Modeling the approximate number system to quantify the contribution of visual stimulus features.* Cognition, 142, 247–265. https ://doi.org/10.1016/j.cognition.2015.05.016
- Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). *CORNET: Modeling the neural mechanisms of core object recognition*. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/408385
- Tarigopula, P., Fairhall, S. L., Bavaresco, A., Truong, N., & Hasson, U. (2023). *Improved prediction of behavioral and neural similarity spaces using pruned DNNs*. Neural Networks, 168, 89–104. https://doi.org/10.1016/j.neunet.2023.08.049