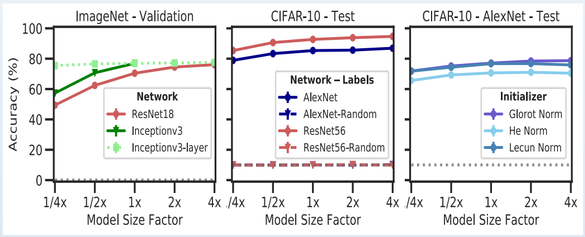
Frivolous Units Help to Explain Non-Overfitting in Overparametrized Deep Neural Networks

Stephen Casper*, Xavier Boix*, Vanessa D'Amario, Ling Guo,
Martin Schrimpf, Kasper Vinken, Gabriel Kreiman
* Shared first authorship

Non-Overfitting Problem

Without explicit regularization, test accuracy does not degrade as network widths increase despite wider networks having a greater capacity for overfitting.



Main Finding

We identify two distinct types of "frivolous" units which constrain complexity: ***prunable*** *and* **redundant** units. Overparameterized deep neural networks consistently constrain their complexity via these units. ArXiv

Implications

Non-overfitting: Overparametrized networks constrain complexity via frivolous units.

Interpretability: Frivolous units may be difficult to interpret.

Compression: Algorithms which target multiple compressible motifs are needed.

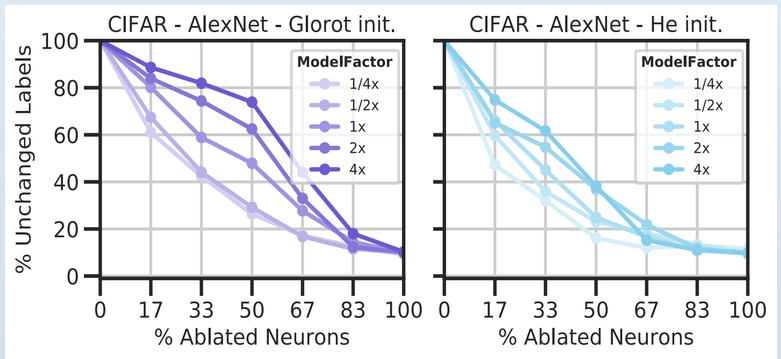
Unit-Level Autoregularization

Prunable units can be dropped out of a network with little effect on the outputs.

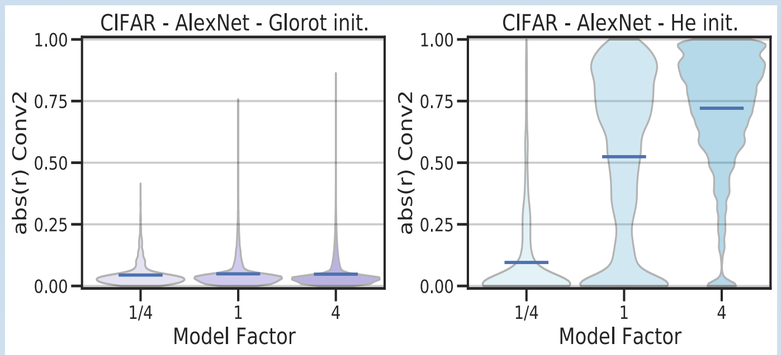
Redundant unit's activations can be expressed well as a linear combination of other units.

Different networks develop different levels of prunability and redundancy.

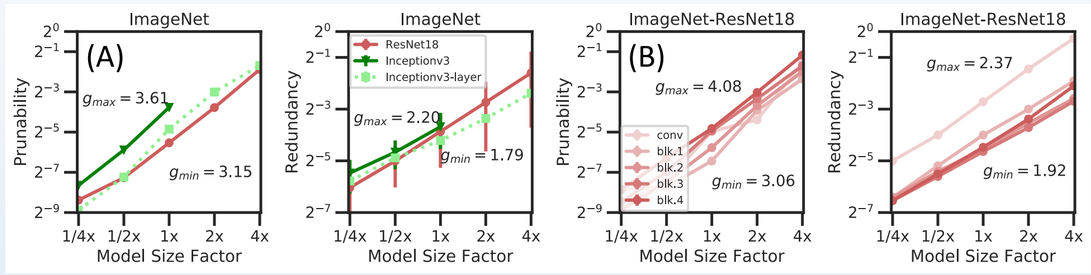
Prunability can be measured by robustness to random ablations.



Redundancy can be measured by linear analysis.



Results

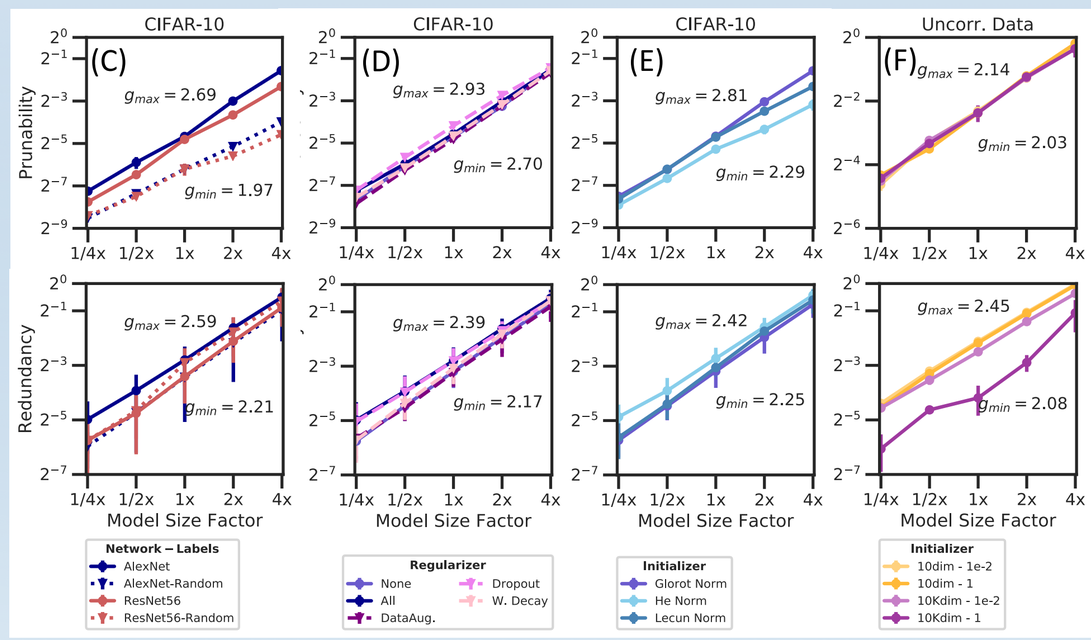


Frivolous units outpace the growth of the network overall: As reflected by the minimum gain factor (g_{min}), as a network's width doubles, frivolous units tend to more than double (all). Gains are similar for individual layers (B).

Frivolous units develop even in networks trained on randomly labeled data: Frivolous units imply capacity constraints, but not generalization (C).

Trends are consistent under explicit regularization: Autoregularization may affect unit-level complexity differently than explicit regularization (D).

Initialization and dataset influence the emergence of frivolous units: Initialization variance and data dimensionality influence these units (E-F).



Acknowledgments: We are grateful to Tomaso Poggio for helpful feedback and discussions. This work is supported by the Center for Brains, Minds and Machines (funded by NSF STC award CCF-1231216), the Harvard office for Undergraduate Research and Fellowships, Fujitsu Laboratories (Contract No. 40008401 and 40008819) and the MIT-Sensetime Alliance on Artificial Intelligence.