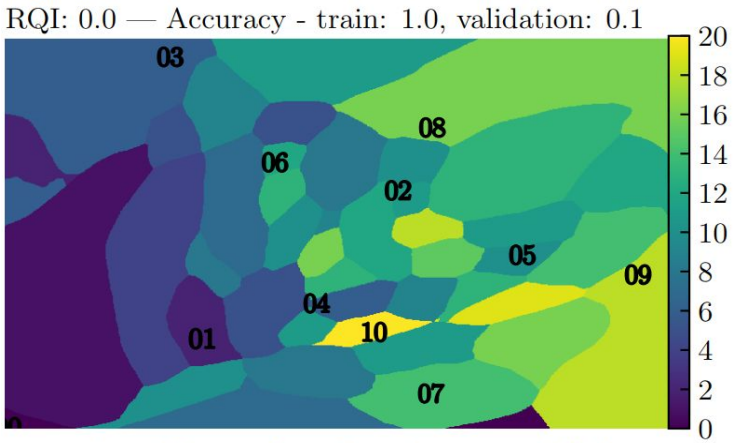
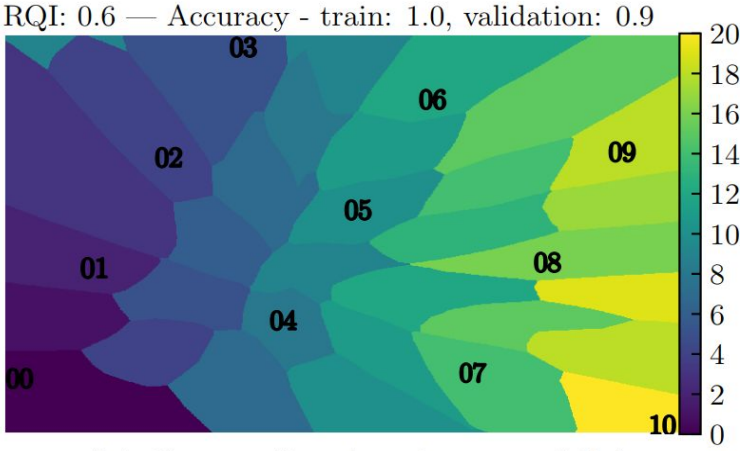
Towards Understanding Grokking: An Effective Theory of Representation Learning

Authors: Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, Mike Williams

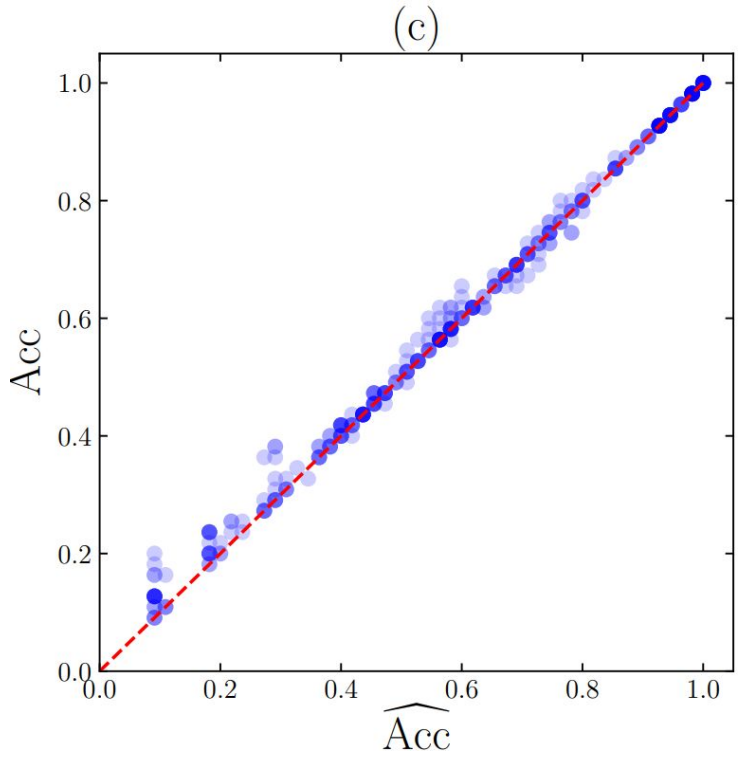
Presenter: Davide Maioli

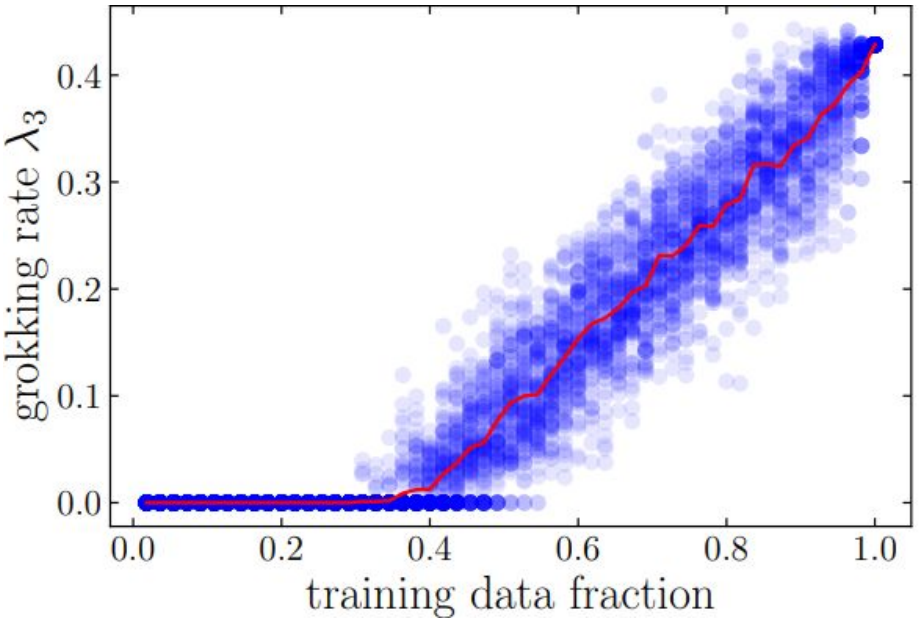


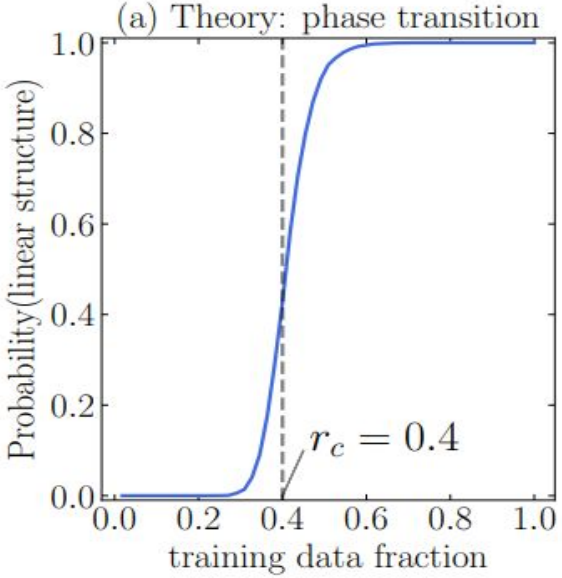
(a) Memorization in toy addition

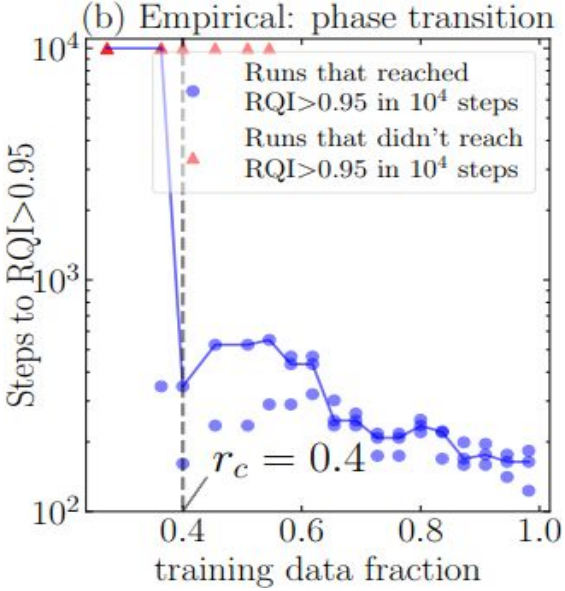


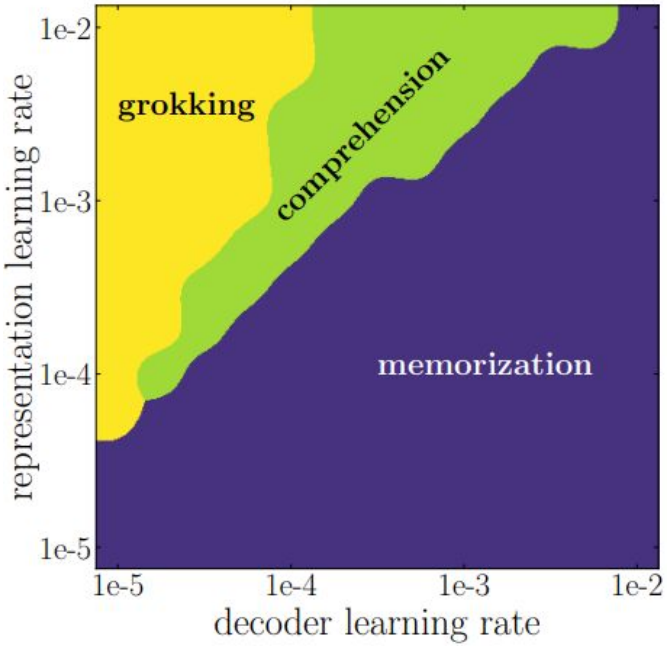
(b) Generalization in toy addition


(c)
1.0
0.8
0.6
0.4
0.2
0.0
Acc
0.0
0.2
0.4
0.6
0.8
1.0
$\widehat{\text{Acc}}$

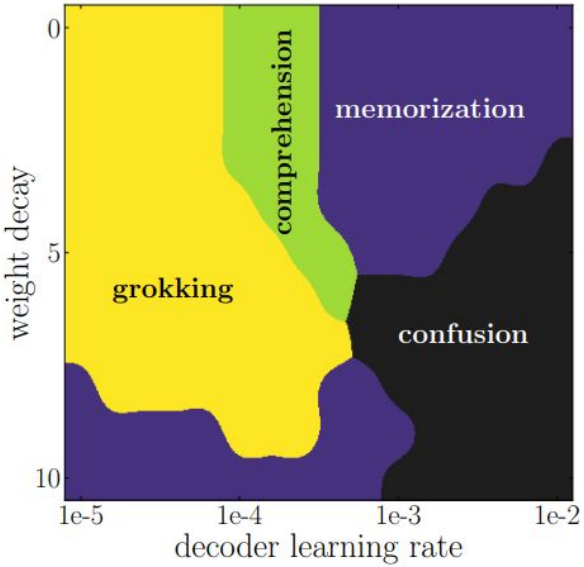

grokking rate λ_3
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
1.0
training data fraction


(a) Theory: phase transition
Probability(linear structure)
training data fraction
$r_c = 0.4$
0.0
0.2
0.4
0.6
0.8
1.0

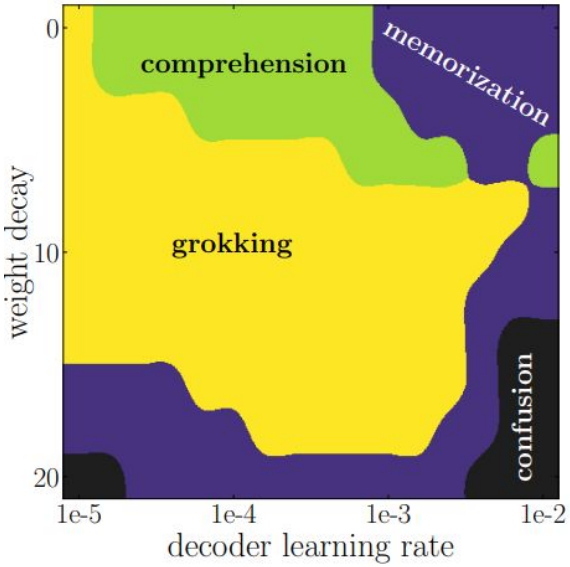

(b) Empirical: phase transition
Steps to RQI>0.95
training data fraction
Runs that reached RQI>0.95 in 10^4 steps
Runs that didn't reach RQI>0.95 in 10^4 steps
$r_c = 0.4$
10^2
10^3
10^4
0.4
0.6
0.8
1.0



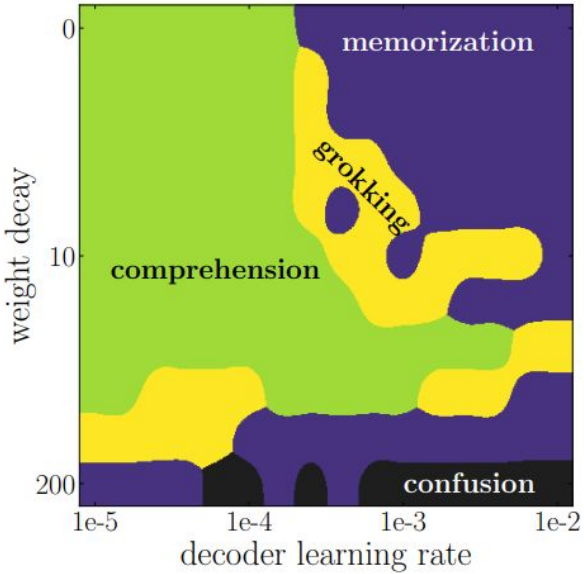
(a) Addition, regression



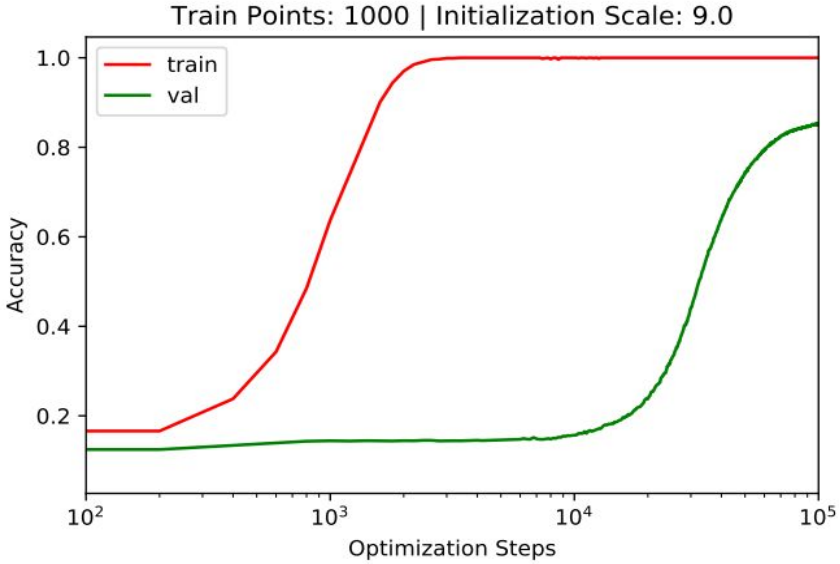
(b) Addition, regression

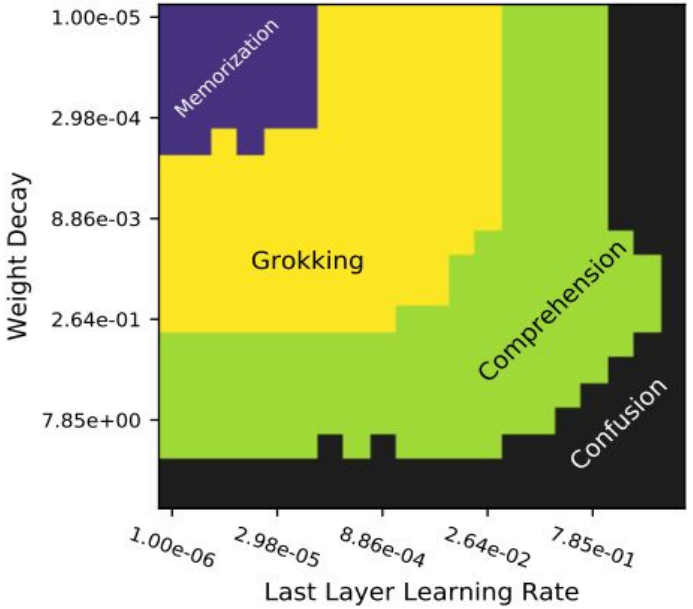


(c) Addition, classification



(d) Permutation, regression


Train Points: 1000 | Initialization Scale: 9.0
train
val
Accuracy
1.0
0.8
0.6
0.4
0.2
10^2
10^3
10^4
10^5
Optimization Steps


Memorization
Grokking
Comprehension
Confusion
Weight Decay
1.00e-05
2.98e-04
8.86e-03
2.64e-01
7.85e+00
1.00e-06
2.98e-05
8.86e-04
2.64e-02
7.85e-01
Last Layer Learning Rate

+ Scientific method

+ Insights about DNN generalization

+ Physics inspired tools

- Broad applicability?