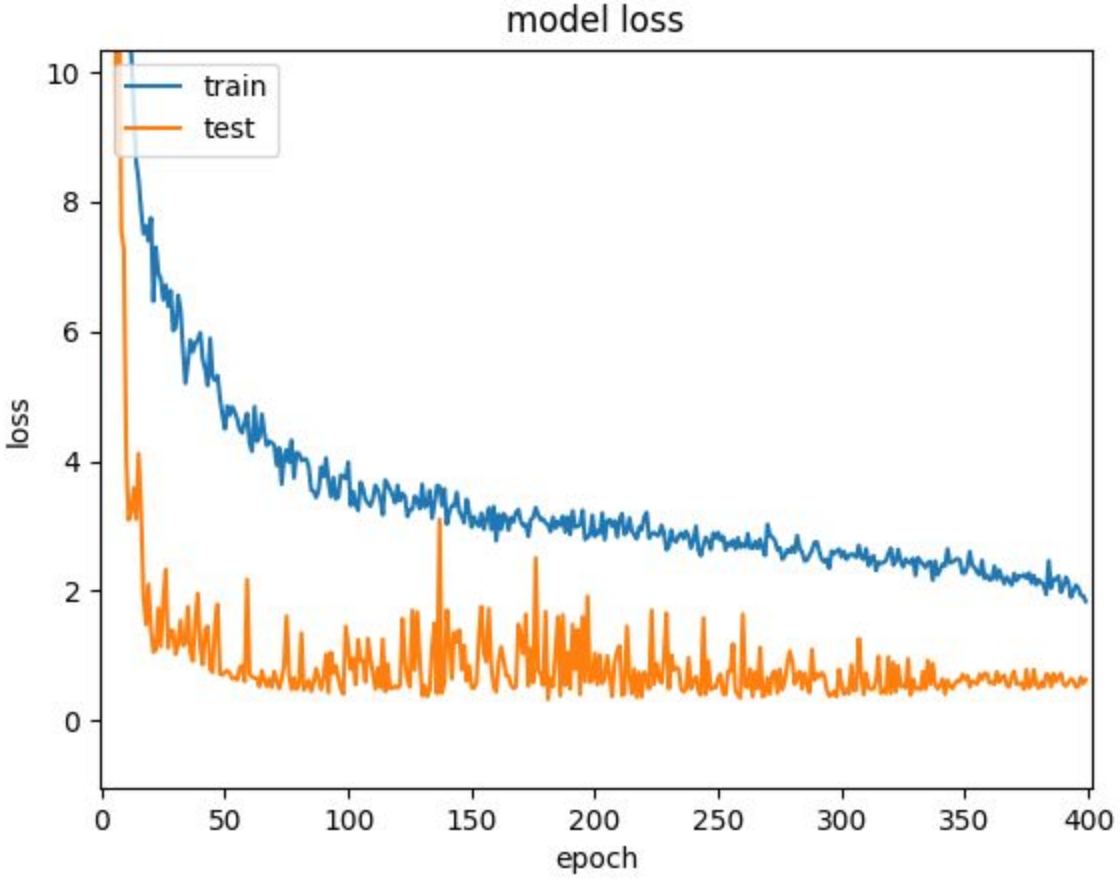
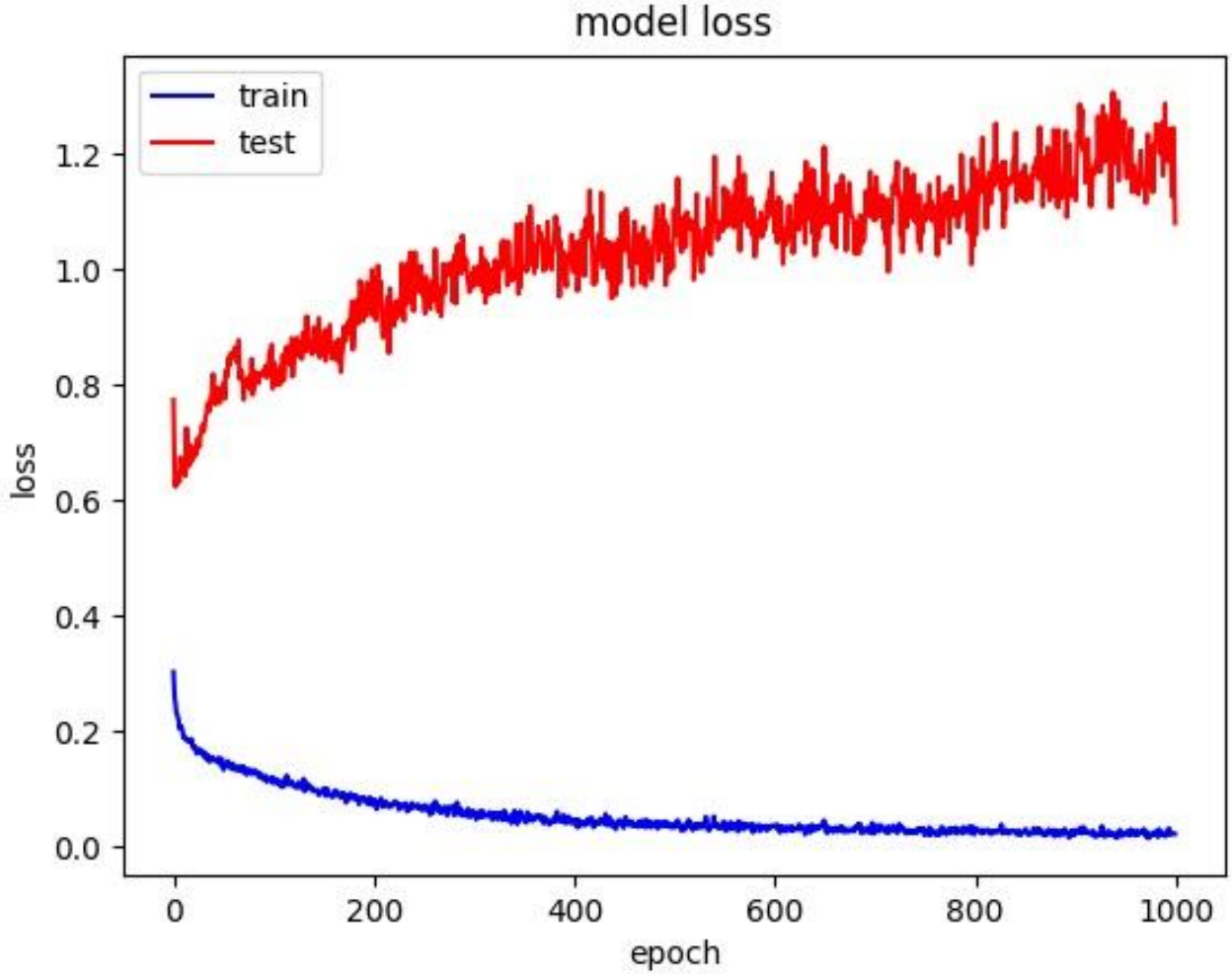
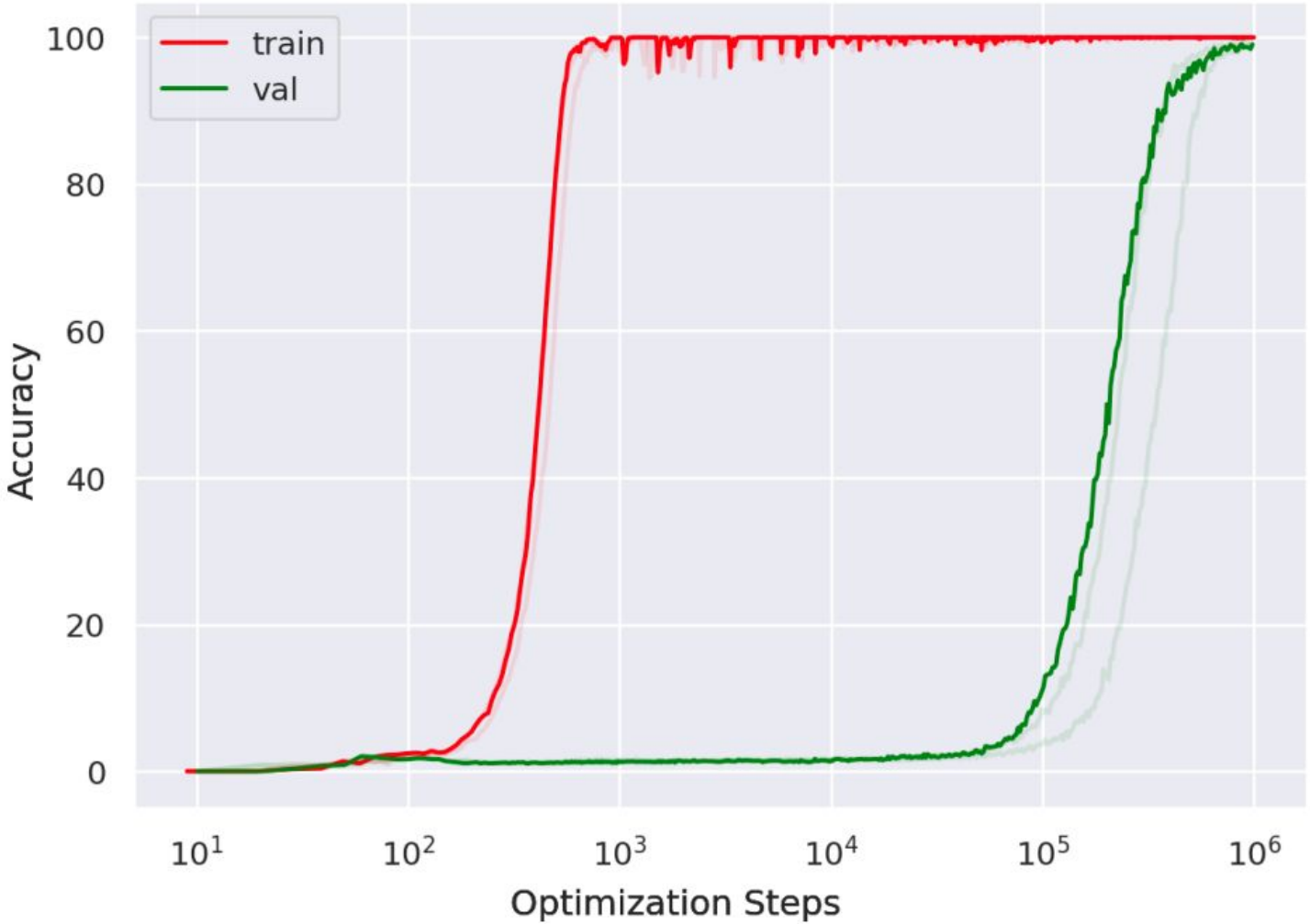
Grokking

Михаил Доманин

Grok – понимать что-либо полностью, буквально на интуитивном уровне, осознавая всю логику и закономерности, а не просто на уровне знания правил


model loss
train
test
loss
epoch
10
8
6
4
2
0
0
50
100
150
200
250
300
350
400

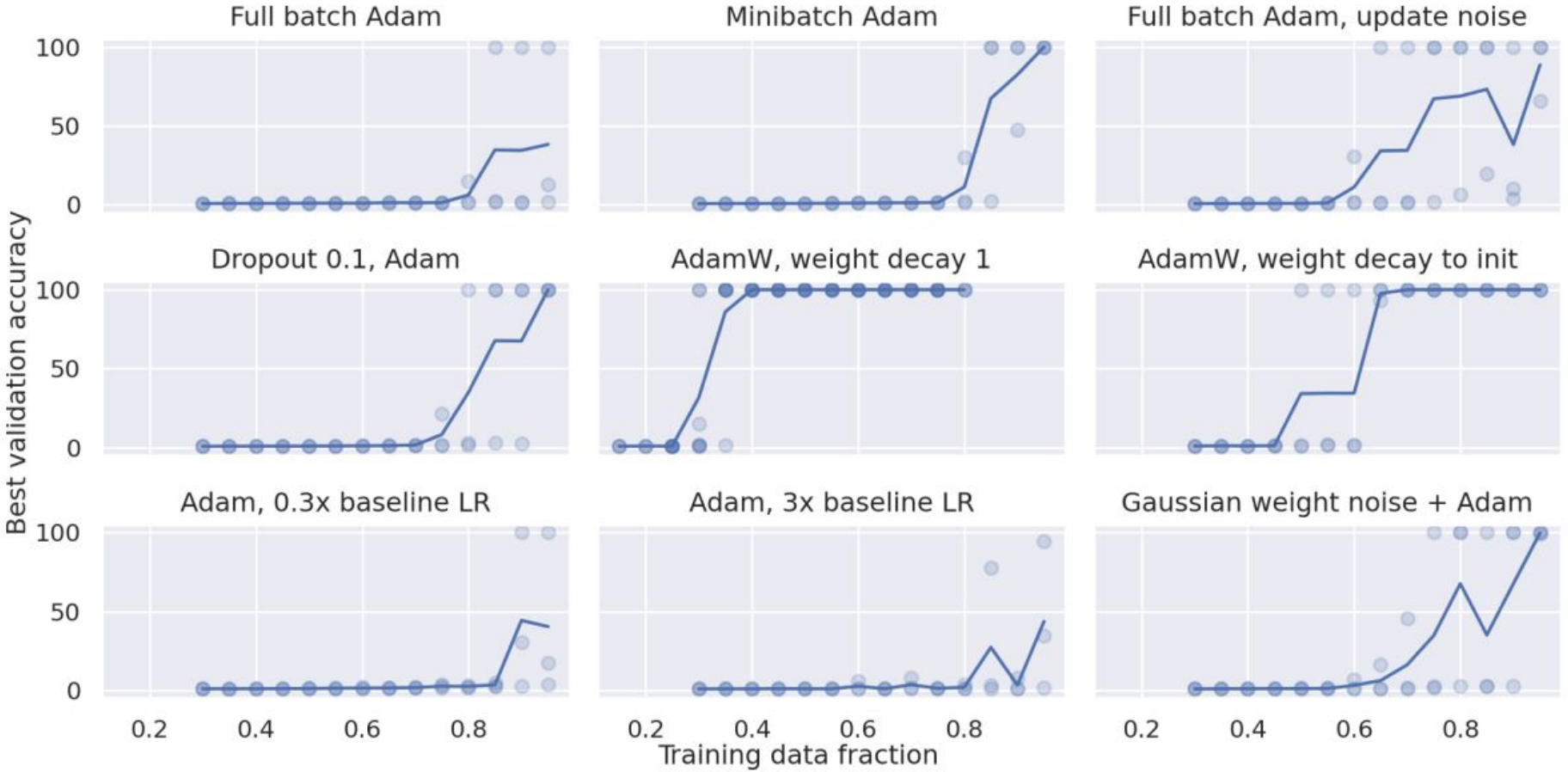

model loss
train
test
loss
epoch
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
200
400
600
800
1000


train
val
100
80
60
40
20
0
Accuracy
10^1
10^2
10^3
10^4
10^5
10^6
Optimization Steps

The following are the binary operations that we have tried (for a prime number $p = 97$):

$$x \circ y = x + y \pmod{p} \text{ for } 0 \leq x, y < p$$
$$x \circ y = x - y \pmod{p} \text{ for } 0 \leq x, y < p$$
$$x \circ y = x/y \pmod{p} \text{ for } 0 \leq x < p, 0 < y < p$$
$$x \circ y = [x/y \pmod{p} \text{ if } y \text{ is odd, otherwise } x - y \pmod{p}] \text{ for } 0 \leq x, y < p$$
$$x \circ y = x^2 + y^2 \pmod{p} \text{ for } 0 \leq x, y < p$$
$$x \circ y = x^2 + xy + y^2 \pmod{p} \text{ for } 0 \leq x, y < p$$
$$x \circ y = x^2 + xy + y^2 + x \pmod{p} \text{ for } 0 \leq x, y < p$$
$$x \circ y = x^3 + xy \pmod{p} \text{ for } 0 \leq x, y < p$$
$$x \circ y = x^3 + xy^2 + y \pmod{p} \text{ for } 0 \leq x, y < p$$
$$x \circ y = x \cdot y \text{ for } x, y \in S_5$$
$$x \circ y = x \cdot y \cdot x^{-1} \text{ for } x, y \in S_5$$
$$x \circ y = x \cdot y \cdot x \text{ for } x, y \in S_5$$


Steps until generalization for product in abstract group S_5
$5 \cdot 10^5$
10^5
10^4
10^3
0.3
0.4
0.5
0.6
0.7
0.8
Training data fraction
Runs that didn't reach 99% val acc in $5 \cdot 10^5$ updates
Runs that reached > 99% val acc in $5 \cdot 10^5$ updates
Median


Full batch Adam
Minibatch Adam
Full batch Adam, update noise
Dropout 0.1, Adam
AdamW, weight decay 1
AdamW, weight decay to init
Adam, 0.3x baseline LR
Adam, 3x baseline LR
Gaussian weight noise + Adam
Best validation accuracy
Training data fraction
0
50
100
0.2
0.4
0.6
0.8

The key feature of the algorithm is calculating $\cos(w(x+y)), \sin(w(x+y))$ with $w = \frac{2\pi}{p}k$ - this is a function of $x+y$ and be mapped to $x+y$, and because $\cos(wx)$ has period $\frac{p}{k}$ we get the $(\mathbf{mod}\ p)$ part for free.

More concretely:

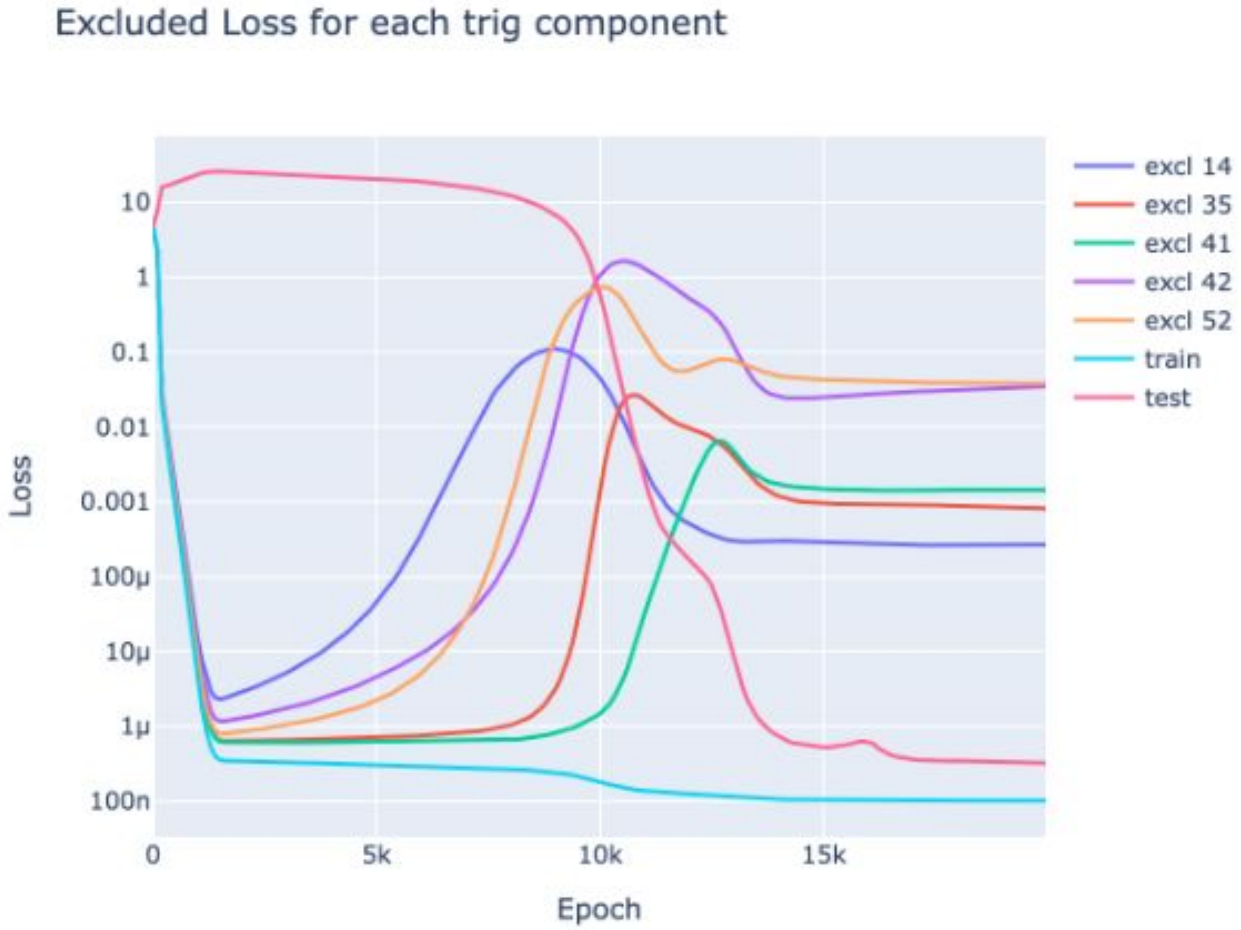
- Inputs x, y are given as one-hot encoded vectors in $\mathbb{R}^p$
- Calculates ° $\cos(wx), \cos(wy), \sin(wx), \sin(wy)$ via a Discrete Fourier Transform (This sounds complex but is just a change of basis on the inputs, and so is just a linear map)
  - $w = \frac{2\pi}{p}k$, k is arbitrary, we just need period dividing p
- Calculates ° $\cos(wx)\cos(wy), \cos(wx)\sin(wy), \sin(wx)\cos(wy), \sin(wx)\sin(wy)$ by multiplying pairs of waves in x and in y
- Calculates ° $\cos(w(x+y)) = \cos(wx)\cos(wy) - \sin(wx)\sin(wy)$ and $\sin(w(x+y)) = \sin(wx)\cos(wy) + \cos(wx)\sin(wy)$ by rearranging and taking differences
- Calculates ° $\cos(w(x+y-z)) = \cos(w(x+y))\cos(wz) + \sin(w(x+y))\sin(wz)$ via a linear map to the output logits z
  - This has an argmax at $z \equiv x+y\ (\mathbf{mod}\ p)$, so post softmax we're done!

Гипотезы

- Grokking возникает в результате длительного блуждания модели по оптимумам функции потерь

Гипотезы

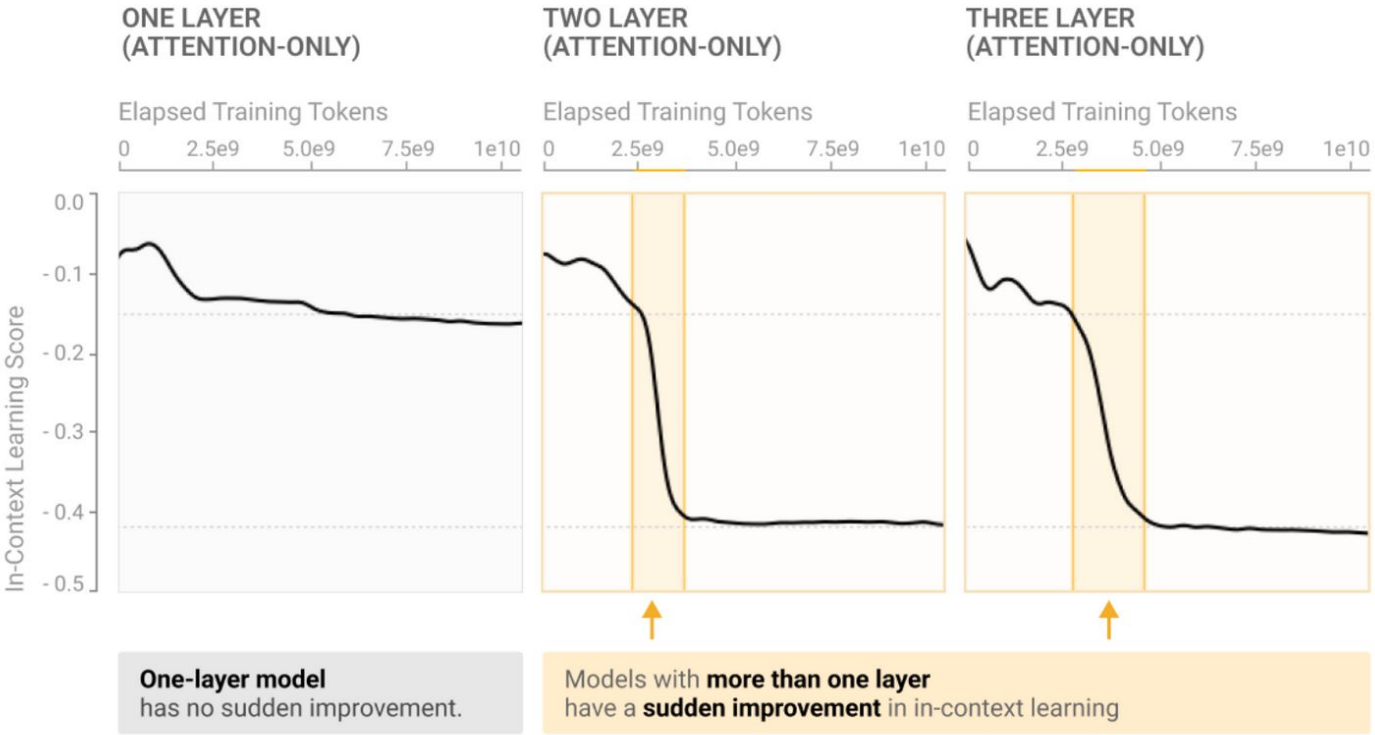
- Grokking возникает в результате длительного блуждания модели по оптимумам функции потерь
- Выход нейросети – комбинация значений разных схем. Некоторые из них потенциально полезны для уменьшения лоса, другие нет. Градиентный спуск “погасит” явно не важные признаки (обнуляя коэффициенты перед ними) и усилит полезные, то есть модель будет постепенно все ближе и ближе к тому, чтобы понять истинную закономерность.


Excluded Loss for each trig component
Loss
10
1
0.1
0.01
0.001
100μ
10μ
1μ
100n
0
5k
10k
15k
Epoch
excl 14
excl 35
excl 41
excl 42
excl 52
train
test

Входные данные

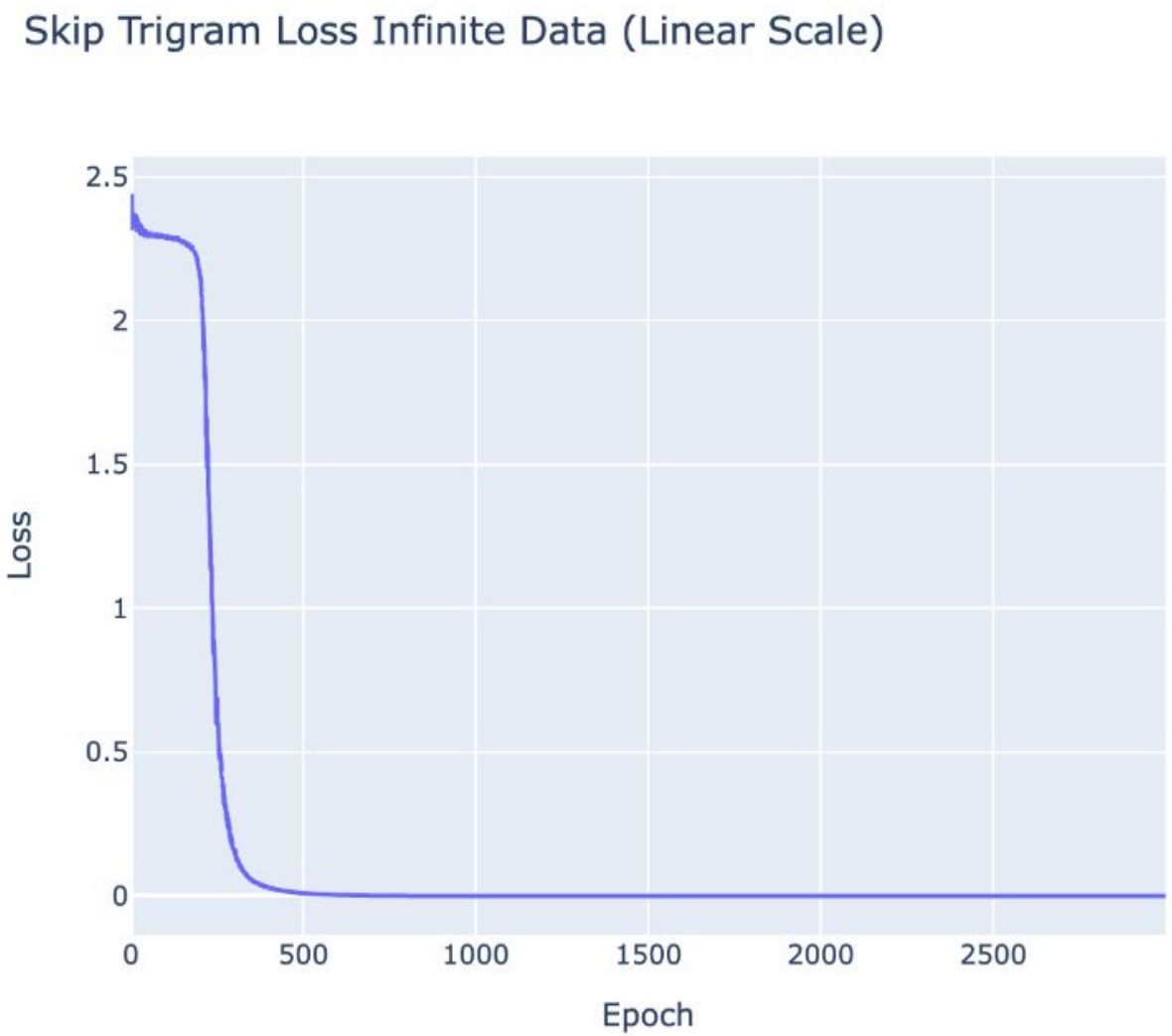
- Данных слишком мало – модель легко запоминает, нет стимула к обобщению
- Данных слишком много – модель легко обобщает
- Данных ни много, ни мало – проявляется Grokking

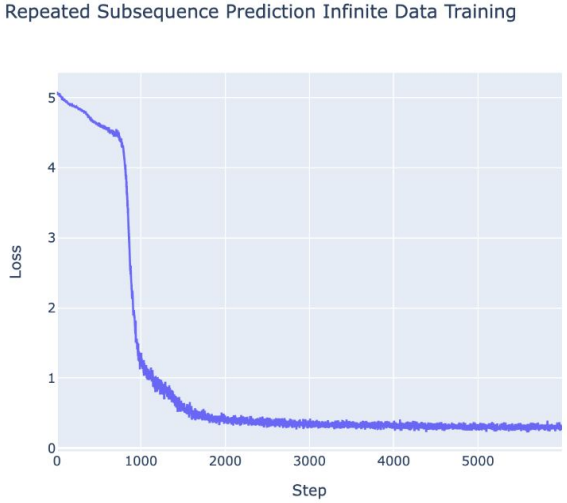
MODELS WITH MORE THAN ONE LAYER HAVE AN ABRUPT IMPROVEMENT IN IN-CONTEXT LEARNING

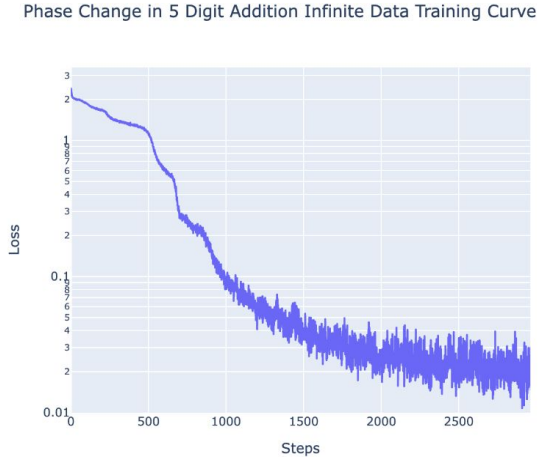


We highlight the **"phase change"** period of training in plots to make visual comparision between plots easier. The highlighted region is selected for each model based on the derivative of in-context learning.

Modular addition mod 113 loss curve, trained on 95% of the data


Skip Trigram Loss Infinite Data (Linear Scale)
Loss
Epoch
2.5
2
1.5
1
0.5
0
0
500
1000
1500
2000
2500


Repeated Subsequence Prediction Infinite Data Training
Loss
Step
5
4
3
2
1
0
0
1000
2000
3000
4000
5000


Phase Change in 5 Digit Addition Infinite Data Training Curve
Loss
Steps
0.1
0.01
0
500
1000
1500
2000
2500

Grokking = Регуляризация + Фазовые изменения + Ограниченные данные

Что насчет реальных задач?

- Grokking полагается на интерпретируемость модели
- Шумы сильно мешают возникновению эффекта
- Сложные модели – комбинация большого количества схем, поиск обобщающего решения вычислительно может быть очень долгим
- Однако…


Per-digit Loss Curves for 5 digit addition (Infinite Data)
Loss
Step
0
0.5
1
1.5
2
2.5
0
200
400
600
800
1000
tok 0
tok 1
tok 2
tok 3
tok 4
tok 5
train_loss