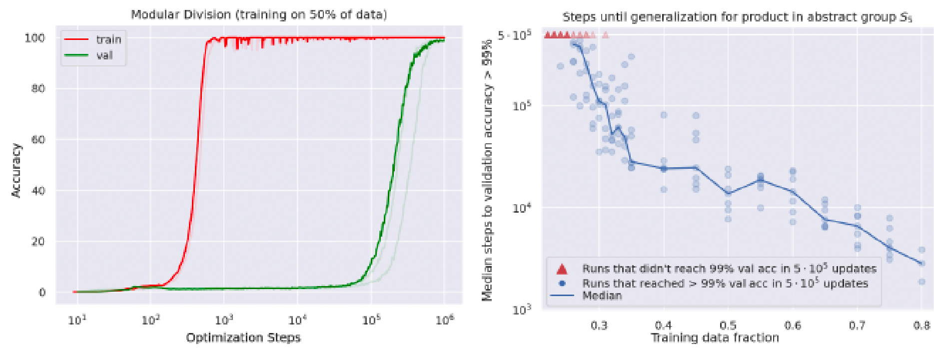
Name: Grokking: Generalization beyond overfitting on small algorithmic datasets

Link: https://mathai-iclr.github.io/papers/papers/MATHAI_29_paper.pdf

A. **Introduction**:
    a. **Grokking**: model transitions from predicting randomly to perfect generalization way beyond the point of overfitting.
    b. Classical statistics considers that the model will not improve once it starts to overfit (**over parameterization**) because it memorizes the training data BUT grokking shows that this is not true empirically in some cases
    c. Grokking is seen on artificial datasets but does not occur easily for natural datasets.
    d. **Contributions**:
        i. Replicable grokking phenomenon under specific conditions
        ii. Median #steps needed to generalize is inversely proportional to the size of training data (provided minimum training data size and optimization budget)
        iii. Show that **weight decay regularization** is influential in making the NN grok.

B. **Description:**
    a. **Aim**: Study generalization of overparameterized NN beyond memorization of finite training dataset
    b. **Data:** Datasets used are small and generated algorithmically. Binary ops like addition, composition of permutations and bivariate polynomials.
    c. **Model**: Decoder part of transformer network (encoder-decoder architecture)
    d. **Double descent of validation loss** is observable in limited cases BUT grokking is observable in a wide **variety of models, optimizers and dataset sizes**
    e. **Generalization performance is measured by validation accuracy.**
    f. **Conclusions:**
        i. **Expected versus Observed**:
            1. **Expected**: Decreasing the amount of training data decreases the generalization performance after convergence
            2. **Observed**: Generalization performance after convergence stays at 100% within a range of training dataset size and optimization budget BUT time (#epochs/ #steps) needed to reach convergence increases drastically with decrease in training data size.



3.

ii. **Variety of operations:**
1. Symmetric operations like x*y, x+y etc are intuitively easier to understand and hence require lesser # of training examples to generalize.
2. Some Complicated operations never generalize even if we use 95% training data.
3. Some operations [x=y (mod p) if y is odd, otherwise x - y (mod p)] that require the model to learn a mix of simple operations can be learnt by the model

iii. Impact of Regularizations:
1. **Weight Decay** was the best in inducing generalization and it reduces the # samples needed by half.
2. **Adding noise to the optimization** process by using mini batches or by adding Gaussian noise to weights before and after computing the gradients also leads to generalization

C. My Thoughts:

a. Interesting observation wrt to generalization beyond overfitting.
b. Closely related to Double descent on Validation loss but not quite the same.
c. Observed only on algorithmically generated small dataset
d. What will happen if we use >2 operands?
e. Why use the S5 abstract set and not anything else?
f. Why does only Weight decay work reasonably well?

References:

- **Deep double descent:** https://arxiv.org/pdf/1912.02292.pdf
- **Weight Decay**:
  - https://arxiv.org/pdf/1711.05101.pdf
  - https://github.com/loshchil/AdamW-and-SGDW
- **Symmetric set:**
  - http://www.efgh.com/math/algebra/permutations.htm
  - https://en.wikipedia.org/wiki/Symmetric_group