Scaling Laws

Deep learning has had recent success due to large increases in access to data and compute. Notably, neural language models such as GPT-3 have over 175 billion parameters (Bussler, 2020). Practical questions arise when aiming to train such a model in a real environment as the financial investment required to train such a model is very high. A common solution to this problem is to train a much smaller model of a similar/same architecture and extrapolate how the accuracy/performance of the model will scale as parameter counts or data counts are increased. This gives a quick litmus test to whether an approach is appropriate for the given task.

Interestingly, in empirical studies (Kaplan et al., 2020) training curves have been observed to follow a power law (Vaswani et al., n.d.) which is roughly independent of model size. This observation has allowed practitioners to use simple statistical models to extrapolate from early segments of the training curve to estimate how future performance will scale with more data. Additionally, it has been noted that the sample efficiency of models increases as parameter sizes are increased. This can be seen in modern large language models.