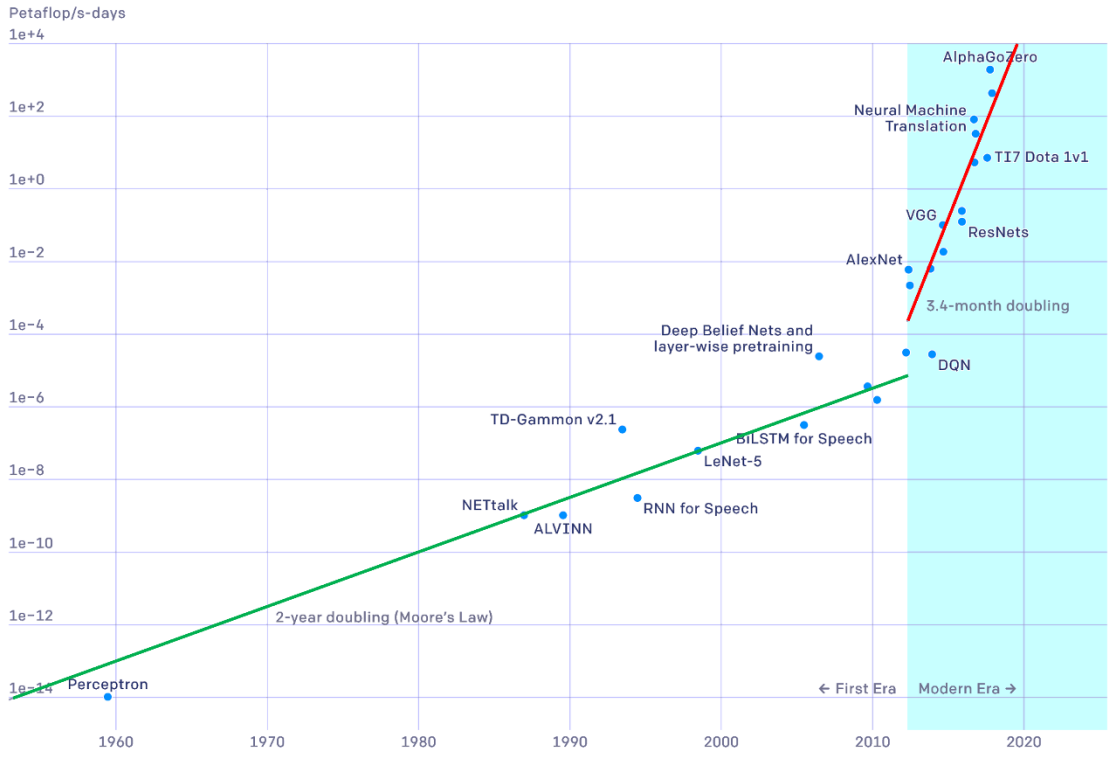

Figure 1.16 Plot of the number of compute cycles, measured in petaflop/s-days, needed to train a state-of-the-art neural network as a function of date, showing two distinct phases of exponential growth. [From OpenAI with permission.]

Figure 1.16 illustrates how the number of compute cycles needed to train a state-of-the-art neural network has grown over the years, showing two distinct phases of growth. The vertical axis has an exponential scale and has units of petaflop/s-days, where a petaflop represents 10^{15} (a thousand trillion) floating point operations, and a petaflop/s is one petaflop per second. One petaflop/s-day represents computation at the rate of a petaflop/s for a period of 24 hours, which is roughly 10^{20} floating point operations, and therefore, the top line of the graph represents an impressive 10^{24} floating point operations. A straight line on the graph represents exponential growth, and we see that from the era of the perceptron up to around 2012, the doubling time was around 2 years, which is consistent with the general growth of computing power as a consequence of Moore's law. From 2012 onward, which marks the era of deep learning, we again see exponential growth but the doubling time is now 3.4 months corresponding to a factor of 10 increase in compute power every year!

It is often found that improvements in performance due to innovations in the architecture or incorporation of more sophisticated forms of inductive bias are soon