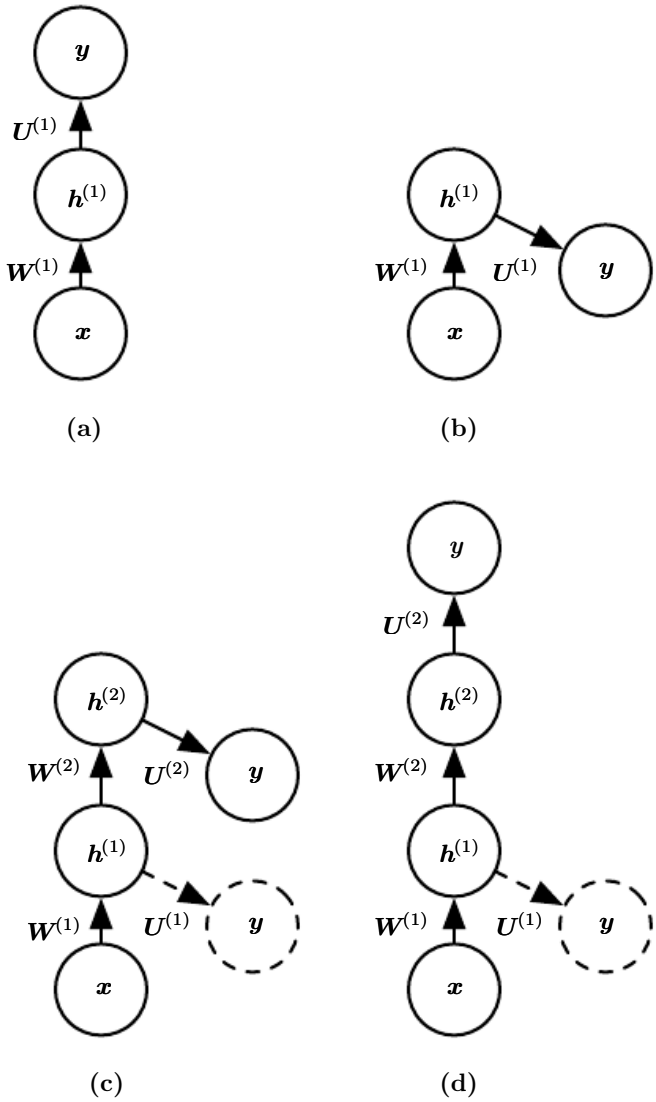

Figure 8.7: Illustration of one form of greedy supervised pretraining (Bengio *et al.*, 2007). *(a)*We start by training a sufficiently shallow architecture. *(b)*Another drawing of the same architecture. *(c)*We keep only the input-to-hidden layer of the original network and discard the hidden-to-output layer. We send the output of the first hidden layer as input to another supervised single hidden layer MLP that is trained with the same objective as the first network was, thus adding a second hidden layer. This can be repeated for as many layers as desired. *(d)*Another drawing of the result, viewed as a feedforward network. To further improve the optimization, we can jointly fine-tune all the layers, either only at the end or at each stage of this process.