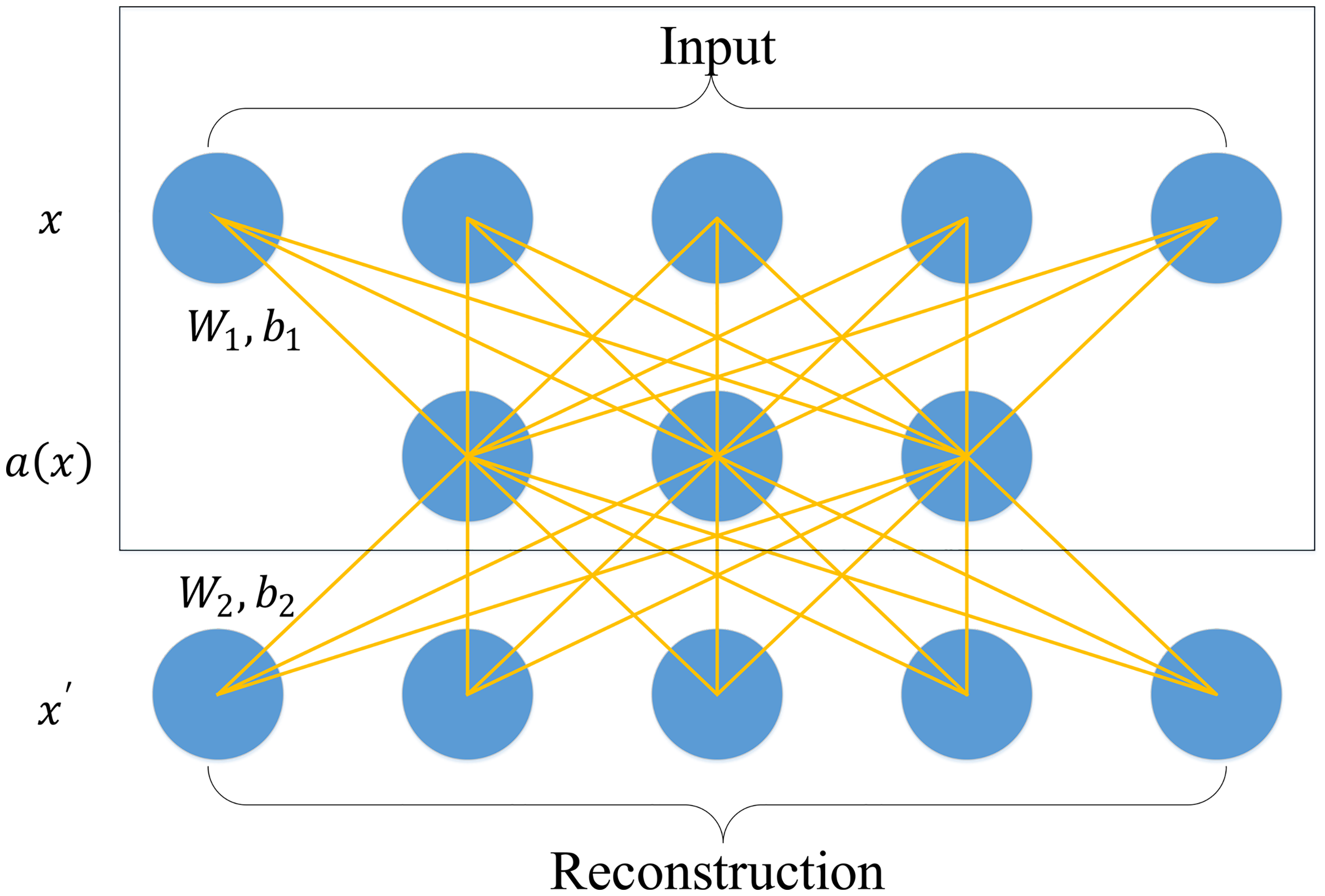
Training a deep network is hard for a lot of reasons, such as vanishing gradients and exploding gradients. Many efforts have been made to solve the problem. One of the most successful architecture is ResNet. The main innovation of ResNets is the use of residual blocks, which include skip connections or shortcuts. These shortcuts allow the network to skip one or more layers during forward and backward propagation, facilitating the flow of gradients through the network. ResNet is extremely useful for super deep networks. However, in our project, due to the limit of data amount, we cannot use very deep network because of the risk of overfitting. ( [Deep Residual Learning for Image Recognition](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=DhtAFkwAAAAJ&citation_for_view=DhtAFkwAAAAJ:ALROH1vI_8AC))

Another important approach to ensure successful training is through careful initialization. The deep network is typically initialized by unsupervised layer-wise training and then fine-tuned by supervised training with labels that can progressively generate more abstract and high-level features layer by layer. In our project, autoencoders is applied for layer-wise training, while SAEs is adopted as the corresponding deep neural network architecture.

Single layer AE is a three-layer neural network; it is illustrated in the following figure.

The first layer and the third layer are the input layer and the reconstruction layer with *k* units, respectively. The second layer is the hidden layer with *n* units, which is designed to generate the deep feature for this single layer AE. The aim of training the single layer AE is to minimize the error between the input vector and the reconstruction vector.

Stacked autoencoders is constructed by stacking a sequence of single-layer AEs layer by layer ([Greedy layer-wise training of deep networks](https://proceedings.neurips.cc/paper/2006/hash/5da713a690c067105aeb2fae32403405-Abstract.html)). The following figure illustrates an instance of an SAE with 5 layers that consists of 4 single-layer autoencoders.