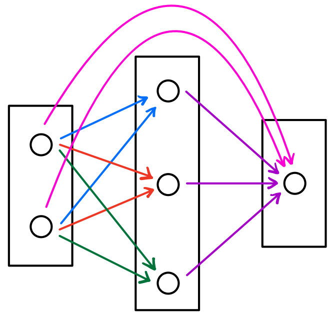
# Lecture 7

## 在这一节中，我们将复习 Residual ANNs (ResNets)：

In this section we review ResNets. Roughly speaking, plain-vanilla feedforward ANNs can be seen as having a computational structure consisting of sequentially chained layers in which each layer feeds information forward to the next layer (cf., for example, Definitions 1.1.3 and 1.3.4 above). ResNets, in turn, are ANNs involving so-called *skip connections* in their computational structure, which allow information from one layer to be fed not only to the next layer, but also to other layers further down the computational structure. In principle, such skip connections can be employed in combinations with other ANN architecture elements, such as fully-connected layers (cf., for instance, Sections 1.1 and 1.3 above), convolutional layers (cf., for example, Section 1.4 above), and recurrent structures (cf., for instance, Section 1.6 below). However, for simplicity we introduce in this section in all mathematical details feedforward fully-connected ResNets in which the skip connection is a learnable linear map (see Definitions 1.5.1 and 1.5.4 below).



ResNets were introduced in He et al. [97] as an attempt to improve the performance of deep ANNs which typically are much harder to train than shallow ANNs (cf., for example, [17, 71, 163]). The ResNets in He et al. [97] only involve skip connections that are identity mappings without trainable parameters, and are thus a special case of the definition of ResNets provided in this section (see Definitions 1.5.1 and 1.5.4 below). The idea of skip connection (sometimes also called *shortcut connections*) has already been introduced before ResNets and has been used in earlier ANN architecture such as the *highway nets* in Srivastava et al. [199, 200] (cf. also [138, 155, 174, 203, 208]). In addition, we refer to [98, 109, 213, 224, 232] for a few successful ANN architecures building on the ResNets in He et al. [97].

## §1 Fully-connected ResNets 的 structured description

1. **Definition:** Fully-connected ResNets 的 structured description (1.5.1)

定义所有 ResNets 组成的集合为：

$$\mathbf{R} = \underbrace{\bigcup_{L\in\mathbb{N}} \underbrace{\bigcup_{l_0,l_1,\ldots,l_L\in\mathbb{N}} \underbrace{\bigcup_{S\subseteq\{(r,k)\in(\mathbb{N}_0)^2:\, r<k\leq L\}}}_{\text{考虑 skip connections 的 possible values}}}_{\text{考虑每层的维度数的 possible values}}}_{\text{考虑 affine function 数 } L \text{ 的 possible values}} \Big(\big(\underbrace{\times_{k=1}^{L}(\mathbb{R}^{l_k\times l_{k-1}}\times\mathbb{R}^{l_k})}_{L\text{组 matrix vector pair 的 possible values}}\big)\times\big(\underbrace{\times_{(r,k)\in S}\mathbb{R}^{l_k\times l_r}}_{\text{skip connections 的 weight matrix 的 possible values}}\big)\Big)$$

2. **Definition:** Fully-connected ResNets (1.5.2)

$\Phi$ 为一个 fully-connected ResNet 当且仅当 $\Phi\in\mathbf{R}$

3. **Lemma:** On an empty set of skip connection (1.5.3)

**Lemma 1.5.3** (On an empty set of skip connections)**.** *Let $L\in\mathbb{N}$, $l_0, l_1, \ldots, l_L\in\mathbb{N}$, $S\subseteq\{(r,k)\in(\mathbb{N}_0)^2 : r<k\leq L\}$. Then*

$$\#\big(\times_{(r,k)\in S}\mathbb{R}^{l_k\times l_r}\big) = \begin{cases} 1 & : S=\emptyset \\ \infty & : S\neq\emptyset. \end{cases} \tag{1.139}$$

若S非空，则有无穷种不同的skip connections

**证明:**

*Proof of Lemma 1.5.3.* Throughout this proof, for all sets $A$ and $B$ let $F(A,B)$ be the set of all functions from $A$ to $B$. Note that

$$\#\big(\times_{(r,k)\in S}\mathbb{R}^{l_k\times l_r}\big) = \#\big\{f\in F\big(S, \cup_{(r,k)\in S}\mathbb{R}^{l_k\times l_r}\big) : (\forall\,(r,k)\in S\colon f(r,k)\in\mathbb{R}^{l_k\times l_r})\big\}. \tag{1.140}$$

This and the fact that for all sets $B$ it holds that $\#(F(\emptyset,B))=1$ ensure that

$$\#\big(\times_{(r,k)\in\emptyset}\mathbb{R}^{l_k\times l_r}\big) = \#(F(\emptyset,\emptyset)) = 1. \tag{1.141}$$

Next note that (1.140) assures that for all $(R, K) \in S$ it holds that

$$\#\left(\times_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}\right) \underset{\text{把 } \times_{(r,k)\in S} \text{ 替换为其中一个元素}}{\geq} \#\left(F(\{(R,K)\}, \mathbb{R}^{l_K \times l_R})\right) = \infty. \tag{1.142}$$

Combining this and (1.141) establishes (1.139). The proof of Lemma 1.5.3 is thus complete. □

4. **Definition:** Fully-connected ResNets 的 realizations (1.5.4)

令 ① 除去 input layer 后的层数 (运算的层数): $L \in \mathbb{N}$

② 各个 layer 的 neuron 数: $l_0, l_1, \ldots, l_L \in \mathbb{N}$

③ 存在 skip connections 的层数的集合: $S \subseteq \{(r,k) \in (\mathbb{N}_0)^2 : r < k \leq L\}$

④ Fully-connected ResNets: $\Phi = ((W_k, B_k)_{k\in\{1,2,\ldots,L\}}, (V_{r,k})_{(r,k)\in S})$

$\in \left(\left(\times_{k=1}^{L} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\right) \times \left(\times_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}\right)\right) \subseteq \mathbf{R}$

⑤ Activation function: $a : \mathbb{R} \to \mathbb{R}$

若对于任意一组 $x_0 \in \mathbb{R}^{l_0}, \ldots, x_L \in \mathbb{R}^{l_L}$ satisfying:

$$x_k = \mathfrak{M}_{\underbrace{a \cdot \mathbb{1}_{(0,L)}(k) + \mathrm{id}_{\mathbb{R}} \cdot \mathbb{1}_{\{L\}}(k)}_{k\in(0,L)\text{时取函数 } a;\ k=L \text{ 取函数 } \mathrm{id}_{\mathbb{R}}}, l_k}\left(W_k x_{k-1} + B_k + \textstyle\sum_{r\in\mathbb{N}_0, (r,k)\in S} V_{r,k} x_r\right), \quad \forall k \in \{1, \ldots, L\}$$

我们有

$$(\mathcal{R}_a^{\mathbf{R}}(\Phi))(x_0) = x_L$$

则函数 $\mathcal{R}_a^{\mathbf{R}}(\Phi)$ 被称为 the realization (function) of the fully-connected ResNet $\Phi$ with activation function $a$

e.g. **Example 1.5.6** (Example for Definition 1.5.2). *Let $l_0 = 1$, $l_1 = 1$, $l_2 = 2$, $l_3 = 2$, $l_4 = 1$, $S = \{(0, 4)\}$, let*

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3), (W_4, B_4)) \in \left(\times_{k=1}^{4} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\right) \tag{1.146}$$

*satisfy*

$$W_1 = \begin{pmatrix} 1 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} 0 \end{pmatrix}, \qquad W_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{1.147}$$

$$W_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad W_4 = \begin{pmatrix} 2 & 2 \end{pmatrix}, \qquad \textit{and} \qquad B_4 = \begin{pmatrix} 1 \end{pmatrix}, \tag{1.148}$$

*and let $V = (V_{r,k})_{(r,k)\in S} \in \times_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}$ satisfy*

$$V_{0,4} = \begin{pmatrix} -1 \end{pmatrix}. \tag{1.149}$$

*Then*

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(5) = 28 \tag{1.150}$$

*(cf. Definitions 1.2.4 and 1.5.4).*

*Proof for Example 1.5.6.* Throughout this proof, let $x_0 \in \mathbb{R}^1$, $x_1 \in \mathbb{R}^1$, $x_2 \in \mathbb{R}^2$, $x_3 \in \mathbb{R}^2$, $x_4 \in \mathbb{R}^1$ satisfy for all $k \in \{1, 2, 3, 4\}$ that $x_0 = 5$ and

$$x_k = \mathfrak{M}_{\mathfrak{r}\mathbb{1}_{(0,4)}(k) + \mathrm{id}_{\mathbb{R}} \mathbb{1}_{\{4\}}(k), l_k}\left(W_k x_{k-1} + B_k + \textstyle\sum_{r\in\mathbb{N}_0, (r,k)\in S} V_{r,k} x_r\right). \tag{1.151}$$

Observe that (1.151) assures that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(5) = x_4. \tag{1.152}$$

Next note that (1.151) ensures that

$$x_1 = \mathfrak{M}_{\mathfrak{r},1}(W_1 x_0 + B_1) = \mathfrak{M}_{\mathfrak{r},1}(5), \tag{1.153}$$

$$x_2 = \mathfrak{M}_{\mathfrak{r},2}(W_2 x_1 + B_2) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}(5) + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 5 \\ 11 \end{pmatrix}\right) = \begin{pmatrix} 5 \\ 11 \end{pmatrix}, \tag{1.154}$$

$$x_3 = \mathfrak{M}_{\mathfrak{r},2}(W_3 x_2 + B_3) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 5 \\ 11 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 5 \\ 11 \end{pmatrix}\right) = \begin{pmatrix} 5 \\ 11 \end{pmatrix}, \tag{1.155}$$

$$\begin{aligned} \text{and} \qquad x_4 &= \underset{\mathrm{id}_{\mathbb{R}}}{\mathfrak{M}_{\mathfrak{r},1}}(W_4 x_3 + B_4 + V_{0,4} x_0) \\ &= \underset{\mathrm{id}_{\mathbb{R}}}{\mathfrak{M}_{\mathfrak{r},1}}\left(\begin{pmatrix} 2 & 2 \end{pmatrix}\begin{pmatrix} 5 \\ 11 \end{pmatrix} + \begin{pmatrix} 1 \end{pmatrix} + \begin{pmatrix} -1 \end{pmatrix}\begin{pmatrix} 5 \end{pmatrix}\right) = \underset{\mathrm{id}_{\mathbb{R}}}{\mathfrak{M}_{\mathfrak{r},1}}(28) = 28. \end{aligned} \tag{1.156}$$

This and (1.152) establish (1.150). The proof for Example 1.5.6 is thus complete. □

例 1: *Exercise* 1.5.1. Let $l_0 = 1$, $l_1 = 2$, $l_2 = 3$, $l_3 = 1$, $S = \{(0,3),(1,3)\}$, let

$$\Phi = ((W_1, B_1),(W_2, B_2),(W_3, B_3)) \in \left(\times_{k=1}^{3}(\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\right) \tag{1.157}$$

satisfy

$$W_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \qquad W_2 = \begin{pmatrix} -1 & 2 \\ 3 & -4 \\ -5 & 6 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \tag{1.158}$$

$$W_3 = \begin{pmatrix} -1 & 1 & -1 \end{pmatrix}, \qquad \text{and} \qquad B_3 = \begin{pmatrix} -4 \end{pmatrix}, \tag{1.159}$$

and let $V = (V_{r,k})_{(r,k)\in S} \in \times_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}$ satisfy

$$V_{0,3} = \begin{pmatrix} 1 \end{pmatrix} \qquad \text{and} \qquad V_{1,3} = \begin{pmatrix} 3 & -2 \end{pmatrix}. \tag{1.160}$$

Prove or disprove the following statement: It holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(-1) = 0 \tag{1.161}$$

(cf. Definitions 1.2.4 and 1.5.4).

Let $x_0 = -1$, and

$$x_1 = M_{r,l_1}(W_1 x_0 + B_1) = M_{r,2}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot (-1) + \begin{bmatrix} 3 \\ 4 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$x_2 = M_{r,l_2}(W_2 x_1 + B_2) = M_{r,3}\left(\begin{bmatrix} -1 & 2 \\ 3 & -4 \\ -5 & 6 \end{bmatrix}\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$$

$$x_3 = M_{id_R, l_3}(W_3 x_2 + B_3 + V_{0,3} x_0 + V_{1,3} x_1)$$

$$= M_{id_R,1}\left([-1 \ 1 \ -1]\begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix} + (-4) + 1 \times (-1) + [3, -2]\begin{bmatrix} 2 \\ 2 \end{bmatrix}\right) = -7$$

Then by the definition.

$$(R_r^R(\Phi, V))(-1) = -7$$

5. Definition: Identity matrices (1.5.5)

令 $d \in N$.

则记 $I_d \in R^{d\times d}$ 为 the identity matrix in $R^{d\times d}$

e.g. $I_2 + I^{2,2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$