深度学习 第一次作业

[redacted]

一 反向传播

1.3.1

$y_i = \gamma \hat{x}_i + \beta$，故

$$\frac{\partial y_i}{\partial \gamma} = \hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}},$$

$$\frac{\partial y_i}{\partial \beta} = 1,$$

$$\text{where } \mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i, \sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2.$$

1.3.2

记 Softmax 函数的输入为$\boldsymbol{x}$，输出为$\boldsymbol{y}$，均有b个神经元，则有

$$\boldsymbol{y}^i = \frac{e^{\boldsymbol{x}^i}}{\sum_{j=1}^{b} e^{\boldsymbol{x}^j}}, i = 1, \dots, b$$

则输出第i个分量对输入第k个分量的梯度为

$$\frac{\partial \boldsymbol{y}^i}{\partial \boldsymbol{x}^k} = \frac{\frac{\partial e^{x_i}}{\partial \boldsymbol{x}^k}\sum_{j=1}^{b} e^{\boldsymbol{x}^j} - e^{\boldsymbol{x}^i}\frac{\partial \sum_{j=1}^{b} e^{\boldsymbol{x}^j}}{\partial \boldsymbol{x}^k}}{\left(\sum_{j=1}^{b} e^{\boldsymbol{x}^j}\right)^2}$$

$$= \begin{cases} \dfrac{e^{\boldsymbol{x}^i}\sum_{j=1}^{b} e^{\boldsymbol{x}^j} - e^{\boldsymbol{x}^i}e^{\boldsymbol{x}^k}}{\left(\sum_{j=1}^{b} e^{\boldsymbol{x}^j}\right)^2} = \boldsymbol{y}^k\left(1 - \boldsymbol{y}^i\right), & k = i \\ \dfrac{-e^{\boldsymbol{x}^i}e^{\boldsymbol{x}^k}}{\left(\sum_{j=1}^{b} e^{\boldsymbol{x}^j}\right)^2} = -\boldsymbol{y}^k\boldsymbol{y}^i. & k \neq i \end{cases}$$

则有

$$\frac{\partial \boldsymbol{y}^i}{\partial \boldsymbol{x}^k} = \boldsymbol{y}^k\left(\mathbf{1}\{i = k\} - \boldsymbol{y}^i\right).$$

1.3.3

记网络中各个中间输出如下图所示。则

$$\boldsymbol{z}_{1\mathrm{A}} = \theta_{1\mathrm{A}}\boldsymbol{x} + \boldsymbol{b}_{1\mathrm{A}},$$

$$\boldsymbol{a}_{1\mathrm{A}} = \sin \boldsymbol{z}_{1\mathrm{A}}$$

$$\boldsymbol{a}_{\mathrm{DP}} = \boldsymbol{M} \circ \boldsymbol{a}_{1\mathrm{A}},$$

$$\widehat{\boldsymbol{y}}_{\mathrm{A}} = \theta_{2\mathrm{A}} \boldsymbol{a}_{\mathrm{DP}} + \boldsymbol{b}_{2\mathrm{A}},$$

$$\boldsymbol{a}_{1\mathrm{B}} = \theta_{1\mathrm{B}} \boldsymbol{x},$$

$$\mu = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{a}_{1\mathrm{B}}^{i} ,$$

$$\boldsymbol{z}_{\mathrm{BN}} = \boldsymbol{a}_{1\mathrm{B}} - \mu + \boldsymbol{b}_{1\mathrm{B}},$$

$$\boldsymbol{a}_{\mathrm{BN}} = \mathrm{ReLU}(\boldsymbol{z}_{\mathrm{BN}})$$

$$\boldsymbol{z}_{2\mathrm{B}} = \theta_{2\mathrm{B}}(\widehat{\boldsymbol{y}}_{\mathrm{A}} + \boldsymbol{a}_{\mathrm{BN}}) + \boldsymbol{b}_{2\mathrm{B}}$$

$$\widehat{\boldsymbol{y}}_{\mathrm{B}} = \mathrm{Softmax}(\boldsymbol{z}_{2\mathrm{B}}).$$

损失函数为

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{1}{2} ||(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i} - \boldsymbol{y}_{\mathrm{A}}^{i})||_2^2 - \sum_{k=1}^{b} \boldsymbol{y}_{\mathrm{B},k}^{i} \log \widehat{\boldsymbol{y}}_{\mathrm{B},k}^{i} \right]$$

1.3.4

(1) 计算$\theta_{2\mathrm{B}}, \boldsymbol{b}_{2\mathrm{B}}$。记$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}^{i}$,$\mathcal{L}^{i} = \frac{1}{2} ||(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i} - \boldsymbol{y}_{\mathrm{A}}^{i})||_2^2 - \sum_{k=1}^{b} \boldsymbol{y}_{\mathrm{B},k}^{i} \log \widehat{\boldsymbol{y}}_{\mathrm{B},k}^{i}$。

$$\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}} = (\widehat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}),$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{2\mathrm{B}}} = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}{\partial \boldsymbol{b}_{2\mathrm{B}}^{i}} \right)^{T} \frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}} = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}} = \frac{1}{m} \sum_{i=1}^{m} (\widehat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}) ,$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{2\mathrm{B}}} = \frac{1}{m} \sum_{i=1}^{m} (\widehat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i})(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i} + \boldsymbol{a}_{\mathrm{BN}}^{i})^{T}.$$

(2) 计算$\theta_{1\mathrm{B}}, \boldsymbol{b}_{1\mathrm{B}}$。

$$\frac{\partial \boldsymbol{a}_{\mathrm{BN}}^{i}}{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}} = \mathrm{diag}\big(\mathrm{sgn}(\boldsymbol{z}_{\mathrm{BN}}^{i})\big) ,$$

$$\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{a}_{\mathrm{BN}}^{i}} = \theta_{2\mathrm{B}}^{T} \frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}},$$

$$\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}} = \left(\frac{\partial \boldsymbol{a}_{\mathrm{BN}}^{i}}{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}} \right)^{T} \frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{a}_{\mathrm{BN}}^{i}} = \left(\theta_{2\mathrm{B}}^{T} (\widehat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}) \right) \circ \mathrm{sgn}(\boldsymbol{z}_{\mathrm{BN}}^{i})$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{1\mathrm{B}}} = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}}{\partial \boldsymbol{b}_{1\mathrm{B}}} \right)^{T} \frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}} = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}} = \frac{1}{m} \sum_{i=1}^{m} \left(\theta_{2\mathrm{B}}^{T} (\widehat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}) \right) \circ \mathrm{sgn}(\boldsymbol{z}_{\mathrm{BN}}^{i}) ,$$

$$\frac{\partial \boldsymbol{z}_{\mathrm{BN},k}^{i}}{\partial \boldsymbol{a}_{1\mathrm{B},j}^{i}}=\frac{\partial \boldsymbol{a}_{1\mathrm{B},k}^{i}}{\partial \boldsymbol{a}_{1\mathrm{B},j}^{i}}-\frac{1}{m}\sum_{n=1}^{m}\frac{\partial \boldsymbol{a}_{1\mathrm{B},n}^{i}}{\partial \boldsymbol{a}_{1\mathrm{B},j}^{i}}=\mathbf{1}\{k=j\}-\frac{1}{m},\Longrightarrow \frac{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}}{\partial \boldsymbol{a}_{1\mathrm{B}}^{i}}=I_{t\times t}-\frac{1}{m}\mathbf{1}_{t\times t},$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{1\mathrm{B}}}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{a}_{1\mathrm{B}}^{i}}\left(\boldsymbol{x}^{i}\right)^{T}=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}}{\partial \boldsymbol{a}_{1\mathrm{B}}^{i}}\right)^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{\mathrm{BN}}^{i}}\left(\boldsymbol{x}^{i}\right)^{T}$$

$$=\frac{1}{m}\sum_{i=1}^{m}\left(I_{t\times t}-\frac{1}{m}\mathbf{1}_{t\times t}\right)\left(\left(\theta_{2\mathrm{B}}^{T}\left(\widehat{\boldsymbol{y}}_{\mathrm{B}}^{i}-\boldsymbol{y}_{\mathrm{B}}^{i}\right)\right)\circ \operatorname{sgn}\left(\boldsymbol{z}_{\mathrm{BN}}^{i}\right)\right)\left(\boldsymbol{x}^{i}\right)^{T}.$$

其中，$I_{t\times t}$表示$t\times t$的单位矩阵，$\mathbf{1}_{t\times t}$表示$t\times t$的全 1 矩阵。

(3) 计算$\theta_{2\mathrm{A}},\boldsymbol{b}_{2\mathrm{A}}$。从$\mathcal{L}$出发，存在$\widehat{\boldsymbol{y}}_{\mathrm{A}},\widehat{\boldsymbol{y}}_{\mathrm{B}}$两条反向传播路径。记$\mathcal{L}=\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}^{i}$，$\mathcal{L}^{i}=\mathcal{L}_{\mathrm{A}}^{i}+\mathcal{L}_{\mathrm{B}}^{i},\mathcal{L}_{\mathrm{A}}^{i}=\frac{1}{2}||(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}-\boldsymbol{y}_{\mathrm{A}}^{i})||_{2}^{2},\mathcal{L}_{\mathrm{B}}^{i}=-\sum_{k=1}^{b}\boldsymbol{y}_{\mathrm{B},k}^{i}\log\widehat{\boldsymbol{y}}_{\mathrm{B},k}^{i}$。

$$\frac{\partial \mathcal{L}^{i}}{\partial \widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}}=\frac{\partial \mathcal{L}_{\mathrm{A}}^{i}}{\partial \widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}}+\frac{\partial \mathcal{L}_{\mathrm{B}}^{i}}{\partial \widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}}=\left(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}-\boldsymbol{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}},$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{2\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial \widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}}{\partial \boldsymbol{b}_{2\mathrm{A}}}\right)^{T}\frac{\partial \mathcal{L}^{i}}{\partial \widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial \mathcal{L}^{i}}{\partial \widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}}=\frac{1}{m}\sum_{i=1}^{m}\left(\left(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}-\boldsymbol{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{2\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left(\left(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}-\boldsymbol{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)\left(\boldsymbol{a}_{\mathrm{DP}}^{i}\right)^{T}.$$

(4) 计算$\theta_{1\mathrm{A}},\boldsymbol{b}_{1\mathrm{A}}$。

$$\frac{\partial \boldsymbol{a}_{\mathrm{DP}}^{i}}{\partial \boldsymbol{z}_{1\mathrm{A}}^{i}}=\operatorname{diag}\left(\boldsymbol{M}\circ\cos\left(\boldsymbol{z}_{1\mathrm{A}}^{i}\right)\right),$$

$$\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{a}_{\mathrm{DP}}^{i}}=\theta_{1\mathrm{A}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}},$$

$$\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{1\mathrm{A}}^{i}}=\left(\frac{\partial \boldsymbol{a}_{\mathrm{DP}}^{i}}{\partial \boldsymbol{z}_{1\mathrm{A}}^{i}}\right)^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{a}_{\mathrm{DP}}^{i}}=\left(\theta_{1\mathrm{A}}^{T}\left(\left(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}-\boldsymbol{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)\right)\circ\boldsymbol{M}\circ\cos\left(\boldsymbol{z}_{1\mathrm{A}}^{i}\right),$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{1\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial \boldsymbol{z}_{1\mathrm{A}}^{i}}{\partial \boldsymbol{b}_{1\mathrm{A}}}\right)^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{1\mathrm{A}}^{i}}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{1\mathrm{A}}^{i}}$$

$$=\frac{1}{m}\sum_{i=1}^{m}\left(\theta_{1\mathrm{A}}^{T}\left(\left(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}-\boldsymbol{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)\right)\circ\boldsymbol{M}\circ\cos\left(\boldsymbol{z}_{1\mathrm{A}}^{i}\right),$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{1\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left(\left(\theta_{1\mathrm{A}}^{T}\left(\left(\widehat{\boldsymbol{y}}_{\mathrm{A}}^{i}-\boldsymbol{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)\right)\circ\boldsymbol{M}\circ\cos\left(\boldsymbol{z}_{1\mathrm{A}}^{i}\right)\right)\left(\boldsymbol{x}^{i}\right)^{T}.$$

综上，损失函数对各个层参数的梯度为：

$$\frac{\partial \mathcal{L}}{\partial \theta_{1\mathrm{A}}} = \frac{1}{m}\sum_{i=1}^{m}\left(\left(\theta_{1\mathrm{A}}^{T}\left(\left(\hat{\boldsymbol{y}}_{\mathrm{A}}^{i} - \boldsymbol{y}_{\mathrm{A}}^{i}\right) + \theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)\right)\circ \boldsymbol{M} \circ \cos\left(\boldsymbol{z}_{1\mathrm{A}}^{i}\right)\right)\left(\boldsymbol{x}^{i}\right)^{T},$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{1\mathrm{A}}} = \frac{1}{m}\sum_{i=1}^{m}\left(\theta_{1\mathrm{A}}^{T}\left(\left(\hat{\boldsymbol{y}}_{\mathrm{A}}^{i} - \boldsymbol{y}_{\mathrm{A}}^{i}\right) + \theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)\right)\circ \boldsymbol{M} \circ \cos\left(\boldsymbol{z}_{1\mathrm{A}}^{i}\right),$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{2\mathrm{A}}} = \frac{1}{m}\sum_{i=1}^{m}\left(\left(\hat{\boldsymbol{y}}_{\mathrm{A}}^{i} - \boldsymbol{y}_{\mathrm{A}}^{i}\right) + \theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right)\left(\boldsymbol{a}_{\mathrm{DP}}^{i}\right)^{T},$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{2\mathrm{A}}} = \frac{1}{m}\sum_{i=1}^{m}\left(\left(\hat{\boldsymbol{y}}_{\mathrm{A}}^{i} - \boldsymbol{y}_{\mathrm{A}}^{i}\right) + \theta_{2\mathrm{B}}^{T}\frac{\partial \mathcal{L}^{i}}{\partial \boldsymbol{z}_{2\mathrm{B}}^{i}}\right),$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{1\mathrm{B}}} = \frac{1}{m}\sum_{i=1}^{m}\left(I_{t\times t} - \frac{1}{m}\mathbf{1}_{t\times t}\right)\left(\left(\theta_{2\mathrm{B}}^{T}\left(\hat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}\right)\right)\circ \operatorname{sgn}\left(\boldsymbol{z}_{\mathrm{BN}}^{i}\right)\right)\left(\boldsymbol{x}^{i}\right)^{T},$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{1\mathrm{B}}} = \frac{1}{m}\sum_{i=1}^{m}\left(\theta_{2\mathrm{B}}^{T}\left(\hat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}\right)\right)\circ \operatorname{sgn}\left(\boldsymbol{z}_{\mathrm{BN}}^{i}\right),$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{2\mathrm{B}}} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}\right)\left(\hat{\boldsymbol{y}}_{\mathrm{A}}^{i} + \boldsymbol{a}_{\mathrm{BN}}^{i}\right)^{T},$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_{2\mathrm{B}}} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\boldsymbol{y}}_{\mathrm{B}}^{i} - \boldsymbol{y}_{\mathrm{B}}^{i}\right).$$

二 代码实现

以下代码运行环境为：Ubuntu 20.04，Python 3.8.12，PyTorch 1.8.2。

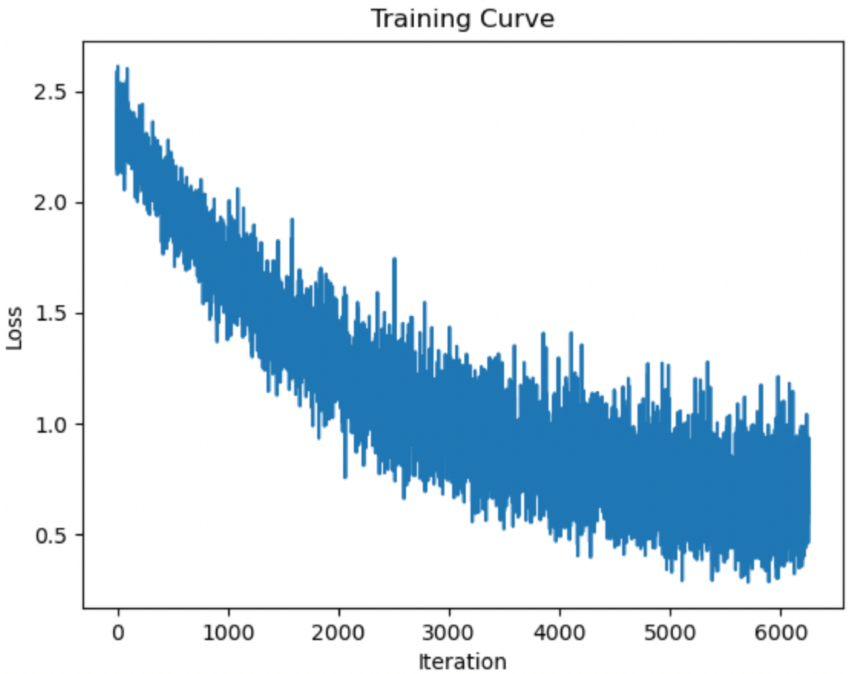
2.1.4.1

将 mlp.py 补全并训练网络，得到输出如下：

```
Hyper-parameters:
Namespace(batch_size=16, epochs=10, hidden_dim=50, lr=0.001)
Dataset information:
training set size: 10000
test set size: 5000
Gradient check of backward propagation:
Relative error of dw2 6.255977816554032e-11
Relative error of db2 4.294765723321833e-12
Relative error of dw1 4.077463739579633e-14
```

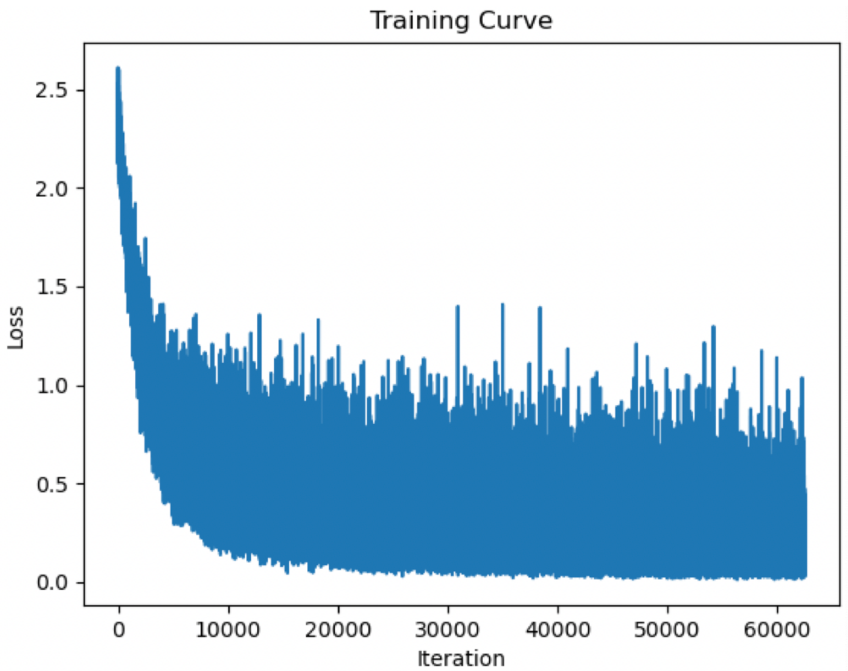
在训练集上准确率为85.04%，在测试集上准确率为83.64%，训练过程中损失函数的变化曲线如下。

```
Relative error of db1 7.970050790502456e-14
If you implement back propagation correctly, all these relative errors should be less than 1e-5.
Start training:
[Epoch #0]
[Iteration #0/625] [Loss #2.320644]
[Iteration #100/625] [Loss #2.446984]
[Iteration #200/625] [Loss #2.223950]
[Iteration #300/625] [Loss #2.024476]
[Iteration #400/625] [Loss #2.108446]
[Iteration #500/625] [Loss #2.186131]
[Iteration #600/625] [Loss #1.950831]
[Epoch #1]
[Iteration #0/625] [Loss #1.797799]
[Iteration #100/625] [Loss #1.666949]
[Iteration #200/625] [Loss #1.747173]
[Iteration #300/625] [Loss #1.612680]
[Iteration #400/625] [Loss #1.396256]
[Iteration #500/625] [Loss #1.667816]
[Iteration #600/625] [Loss #1.544608]
[Epoch #2]
[Iteration #0/625] [Loss #1.658885]
[Iteration #100/625] [Loss #1.608611]
[Iteration #200/625] [Loss #1.589026]
[Iteration #300/625] [Loss #1.506917]
[Iteration #400/625] [Loss #1.164106]
[Iteration #500/625] [Loss #1.355580]
[Iteration #600/625] [Loss #1.441025]
...
[Epoch #9]
[Iteration #0/625] [Loss #0.945138]
[Iteration #100/625] [Loss #0.625108]
[Iteration #200/625] [Loss #0.650663]
[Iteration #300/625] [Loss #0.851720]
[Iteration #400/625] [Loss #0.616638]
[Iteration #500/625] [Loss #0.670256]
[Iteration #600/625] [Loss #0.678758]
Top-1 accuracy on the training set 0.8504
Top-1 accuracy on the test set 0.8364
```

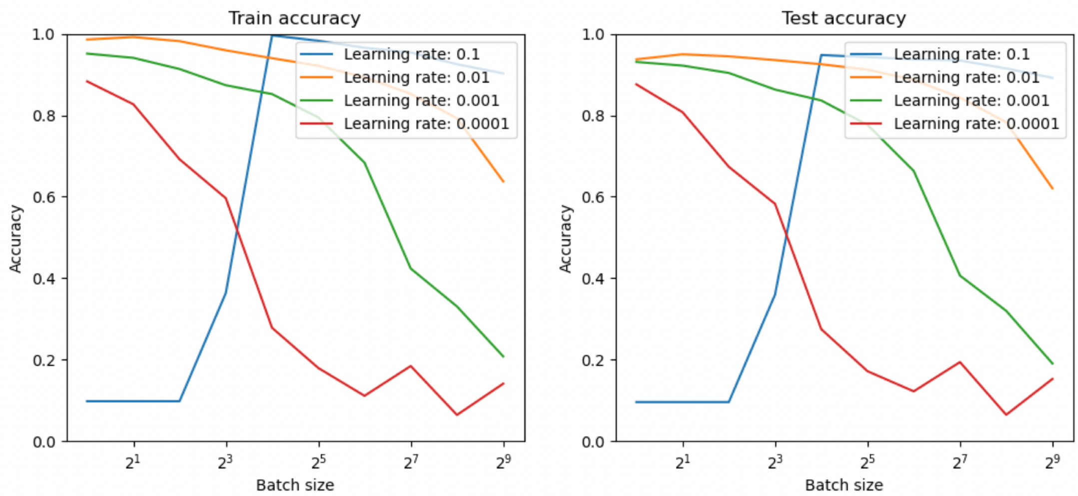


2.1.4.2

由上可以看出，10 个 epochs 后损失函数并未收敛。训练 100 个 epochs，得到在训练集上准确率为 94.38%，在测试集上准确率为 92.50%，训练过程中损失函数的变化曲线如下。



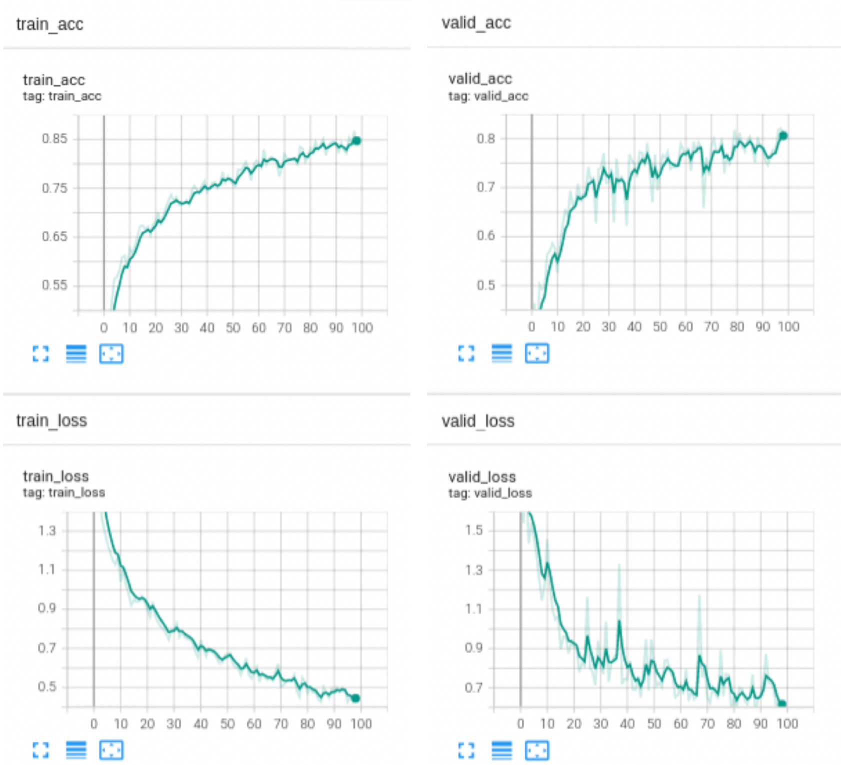
调节学习率和 batch size，可以获得更好准确率。固定训练 10 个 epochs，得到在训练集和测试集上的准确率如下所示。



学习率过低会导致收敛速度减缓，从而准确率较低。而高学习率、batch size 小时，由于噪声过大，损失函数也难以收敛。最好情况下，取学习率为 0.01，batch size 为 2，可得训练集准确率 99.20%，测试集准确率 94.96%。

2.2.3.1

直接运行 main.py，训练 epochs 数为 100，得到训练集最高准确率为 84.20%，验证集最高准确率 81.24%。每个 epoch 的训练集、验证集的准确率、损失函数变化情况如下所示。由曲线可知损失函数并未完全收敛，可继续训练或使用学习率策略达到更高的准确率。



2.2.3.2

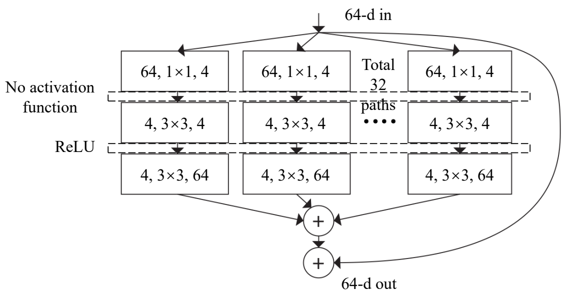
参考 PyTorch 的 resnet.py 源码，参考 ResNeXt [1]，并通过一定的尝试调整，最终设计了模型 B。设计了多个 Pathway，以达到增宽网络、增加拟合能力的目的，并且参数量不至过大。另外有如下考虑与调整：

(1) 由于本任务训练集较小，过大的网络易导致过拟合，因此调整架构时不增加网络深度。并且由于分类输出共 10 类，远小于 ImageNet 数据集的 1000 类，故适当减小中间层的通道数，使网络尽快收敛。

(2) 参考 ConvNeXT [2]的描述，Pathway 的宽度较大可取得较大准确率提升，同时可使用 Inverted Bottleneck 取代 Bottleneck。因此实现中使用 Inverted Bottleneck 作为基本单元，中间 3×3 隐藏层通道数比输入、输出多，以获得更好的拟合能力。

(3) 进一步参考 ConvNeXT [2]所描述的调节方法，将部分激活函数替换为线性激活，由此每个 Block 内只有一个 ReLU 激活层。这样可获得更好的收敛与准确率。

由此，设计模型 B 的单个 Block 结构如下图(a)所示，整个网络的参数配置以及与模型 A 的对比如下图(b)所示。由于使用 Pathway，即分组卷积，模型的总参数量大幅减少。

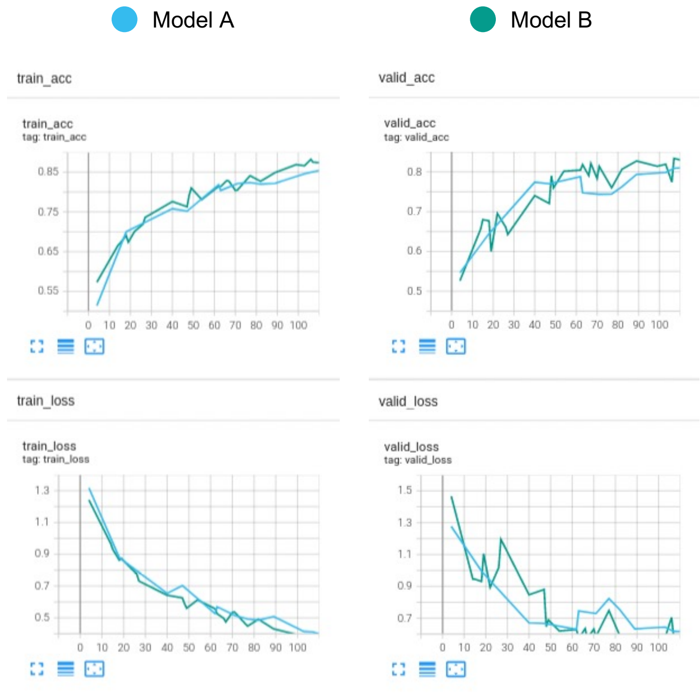


(a)

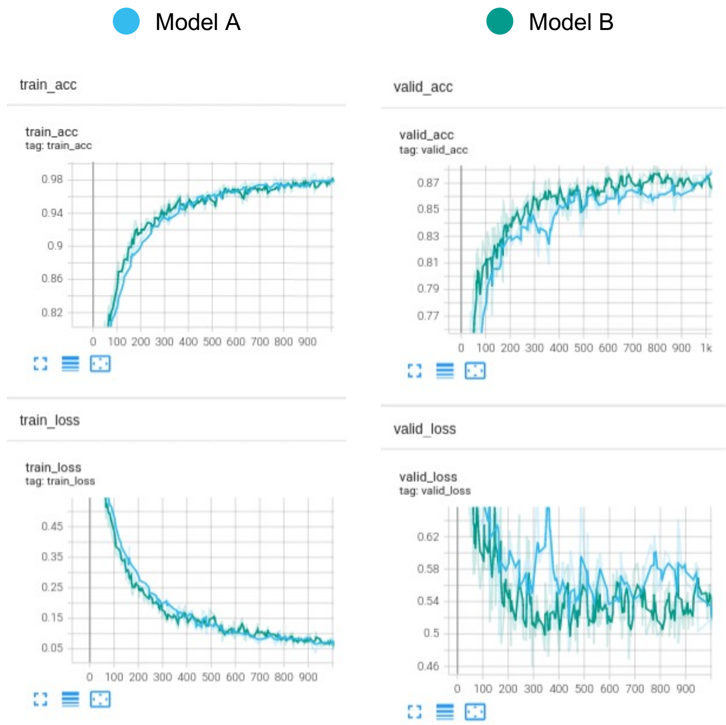
Stage	Output	Model A	**Model B**
conv1	112×112	7×7, 16, stride 2	7×7, 16, stride 2
conv2	56×56	3×3 max pool, stride 2	3×3 max pool, stride 2
		$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times2$	$\begin{bmatrix}1\times1, 128\\3\times3, 128, C=32\\1\times1, 64\end{bmatrix}\times2$
conv3	28×28	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times2$	$\begin{bmatrix}1\times1, 256\\3\times3, 256, C=32\\1\times1, 128\end{bmatrix}\times2$
conv4	14×14	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times2$	$\begin{bmatrix}1\times1, 512\\3\times3, 512, C=32\\1\times1, 256\end{bmatrix}\times2$
conv5	7×7	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times2$	$\begin{bmatrix}1\times1, 1024\\3\times3, 1024, C=32\\1\times1, 512\end{bmatrix}\times2$
	1×1	global average pool 10-d fc, softmax	global average pool 10-d fc, softmax

(b)

训练 epochs 数为 100，得到模型 B 在训练集最高准确率为 86.90%，验证集最高准确率 82.72%。每个 epoch 的训练集、验证集的准确率、损失函数变化情况以及与模型 A 的对比如下所示，其中已把曲线平滑关闭。可见由于波动幅度较大，并且模型仍未完全收敛，模型 B 与模型 A 的性能差异并不明显。

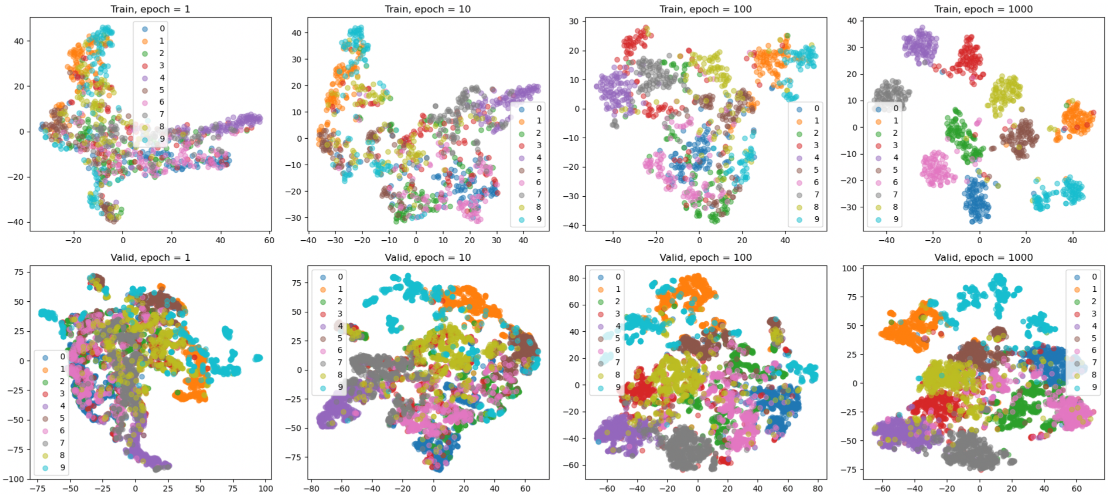


进一步，训练至 1000 个 epoch，得到模型 B 在训练集最高准确率为 98.40%，验证集最高准确率 87.69%。使用系数为 0.6 的平滑后，得到每个 epoch 的训练集、验证集的准确率、损失函数变化情况以及与模型 A 的对比如下所示。可见模型 B 相较模型 A 收敛较快，相同 epoch 下模型 B 在训练集（训练早期）、验证集上平均比模型 A 高约 1% – 2%，在训练后期模型 B 与模型 A 在训练集上趋同，均接近于 98%。并且，模型 B 的收敛较模型 A 更为平滑。



2.2.3.3

对不同 epoch 下模型 A 最后全连接层前的特征进行 t-SNE 降维并可视化，使用图片的真实标签进行标记，得到结果如下。



可见，在训练集上 10 个类别逐渐分离开，至第 1000 个 epoch 时已经可以较为清晰地分辨，仅有少数样本有较大的偏离，此时对应准确率已达到 98%左右。而对于验证集，虽然 10 个类别也是逐渐分离开，但之间间隔较小，并且部分类别与其他类交错分布，也存在较多偏离聚类中心的点，此时对应准确率为 87%左右。可见模型对数据集有较强的过拟合现象。

2.2.3.4

(1) 数据增强。此部分均训练 1000 个 epochs，无学习率调整策略。

代码中已有随机水平翻转、随机裁剪两种数据增强。进一步独立地添加如下几种变换：

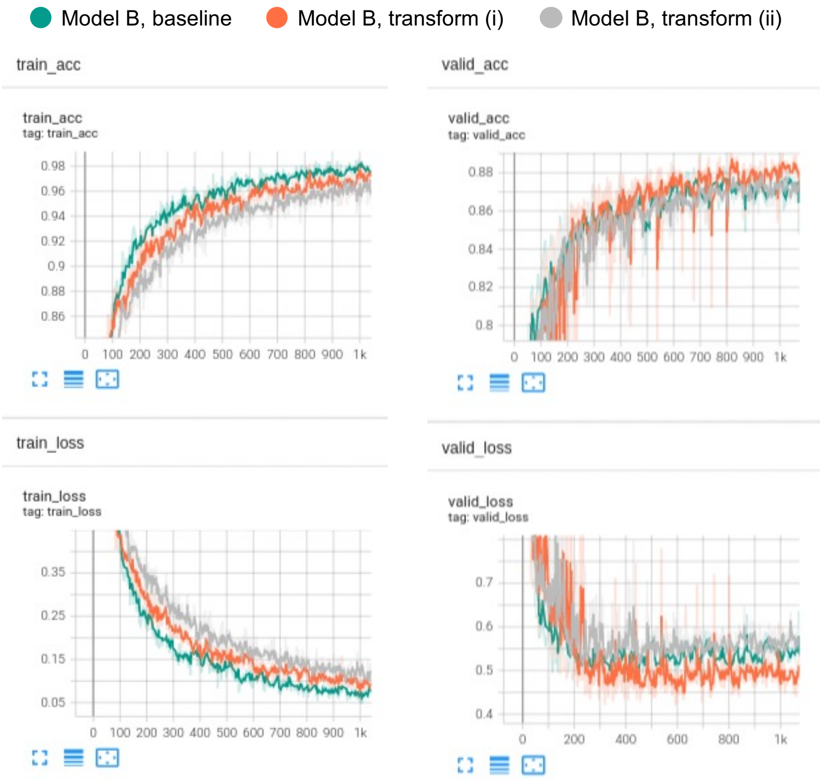
(i) 随机纵向翻转+随机 90 度旋转。与已有的随机水平翻转结合，可遍历完整的[0, 90, 180, 270]度旋转、翻转相组合的共 8 种变换模式。

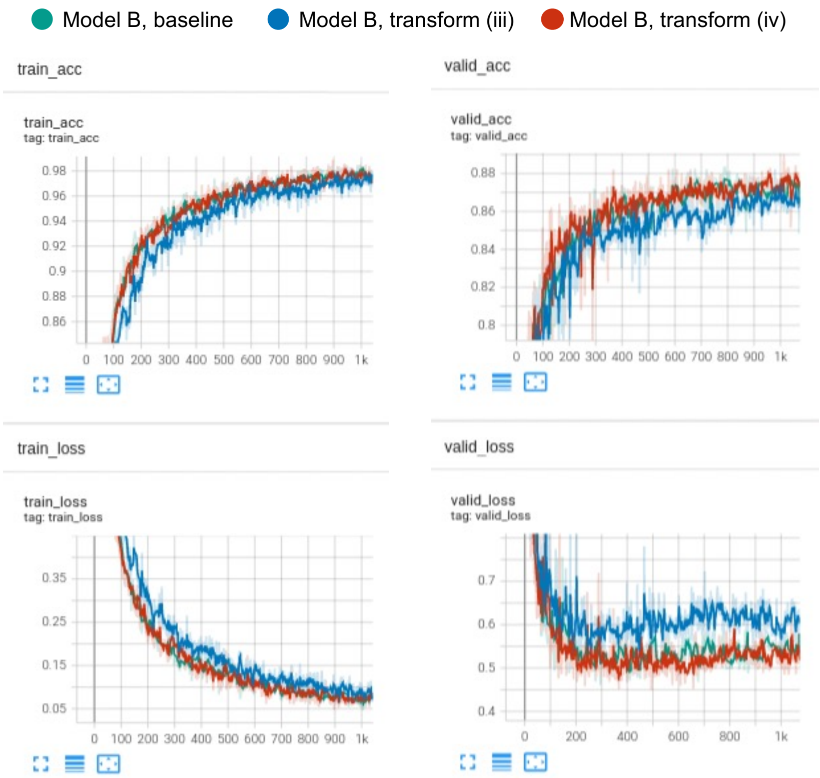
(ii) 随机任意角度旋转。与已有的随机水平翻转结合，可遍历任意角度旋转与翻转相结合的变换模式。

(iii) 随机擦除。随机选取图片一块区域并删除。

(iv) 加入随机噪声。在图像归一化均值、方差后加入标准差为 0.02 的高斯噪声，约为图像标准差的 1/10。

得到训练、验证曲线与原来方法的对比如下。





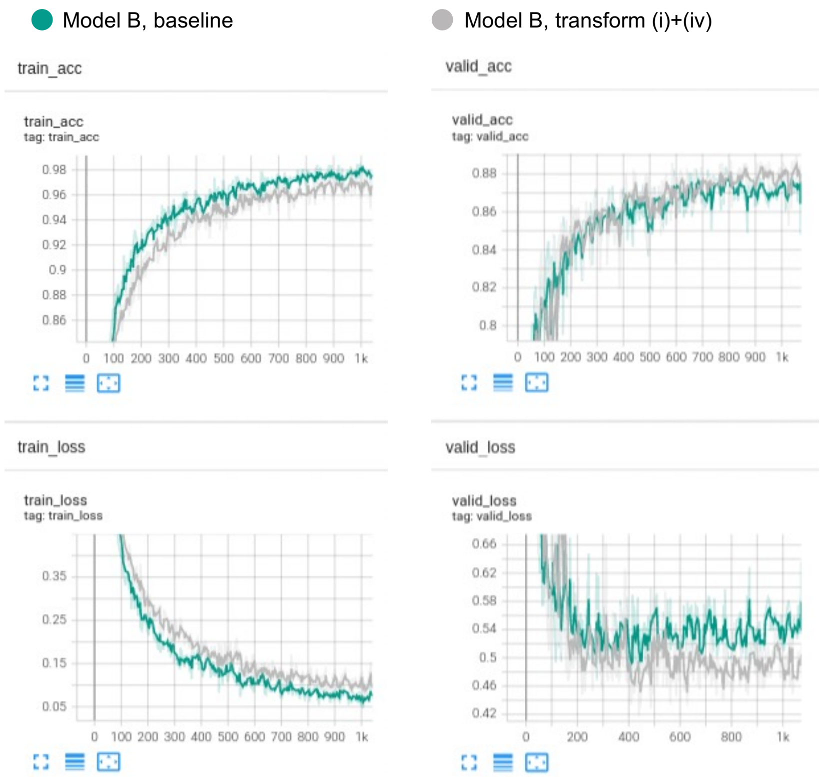
总结四种方式在训练集、验证集上的准确率如下：

	原方法	(i)	(ii)	(iii)	(iv)
训练集准确率	98.40%	97.90%	95.90%	97.40%	97.80%

验证集准确率	87.69%	88.39%	86.94%	86.61%	87.69%

总结学习曲线以及准确率对比可见，(i)、(iv)是较为有效的方法，在验证集上体现出一定优势。这可能是由于(ii)、(iii)中，涉及到图像裁剪或旋转后空的位置需要补全，可能会导致遥感图像信息的一定错误。

由此，组合选用变换(i)+(iv)，得到训练曲线如下。单独增加整 90 度随机旋转及随机翻转会导致增加不同难度的 batch，使得训练前期验证集上出现较大波动；单独为图像增加高斯噪声，会平衡各个 batch 的难度，使得曲线波动较为平缓；将两者结合，可较为稳健地收敛到更好性能。最终选用变换(i)+(iv)后训练集准确率为 97.40%，验证集准确率为 88.96%。观察学习曲线，可较为明显地看到，虽然训练集上准确率下降了约 1% – 2%，但验证集上准确率上升了约 1%。增加数据增强，相当于增加了正则化项，使得网络过拟合程度有所减轻。



(2) 学习率策略。此部分均训练 3000 个 epochs，使用原本的图像增强策略，未说明参数均取默认值（例如(a)(b)中，学习率每次调整为原先的 10%）。

代码中学习率为恒定值。初始学习率定为 0.001，分别使用如下学习率策略，以 epoch 为学习率变化的最小单位，其中参数经过大致实验可取得较优效果：

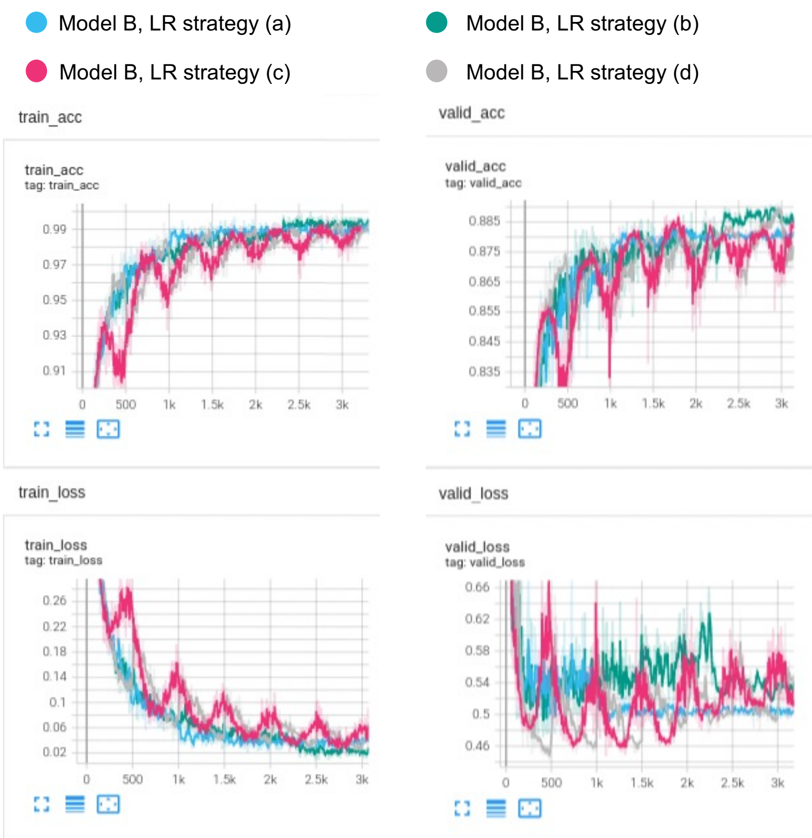
(a) 固定周期减少学习率(StepLR)。具体地，取步长为 1000 个 epochs。

(b) 平坦期减少学习率(ReduceLROnPlateau)。具体地，取容忍时间为 500 个 epochs。

(c) 余弦学习率退火(CosineAnnealingLR)。具体地，取周期为 250 个 epochs。

(d) 余弦学习率退火重启(CosineAnnealingWarmRestarts)。具体地，取周期为 500 个 epochs。

得到训练、验证曲线与原来方法的对比如下。



总结四种方式在训练集、验证集上的准确率如下：

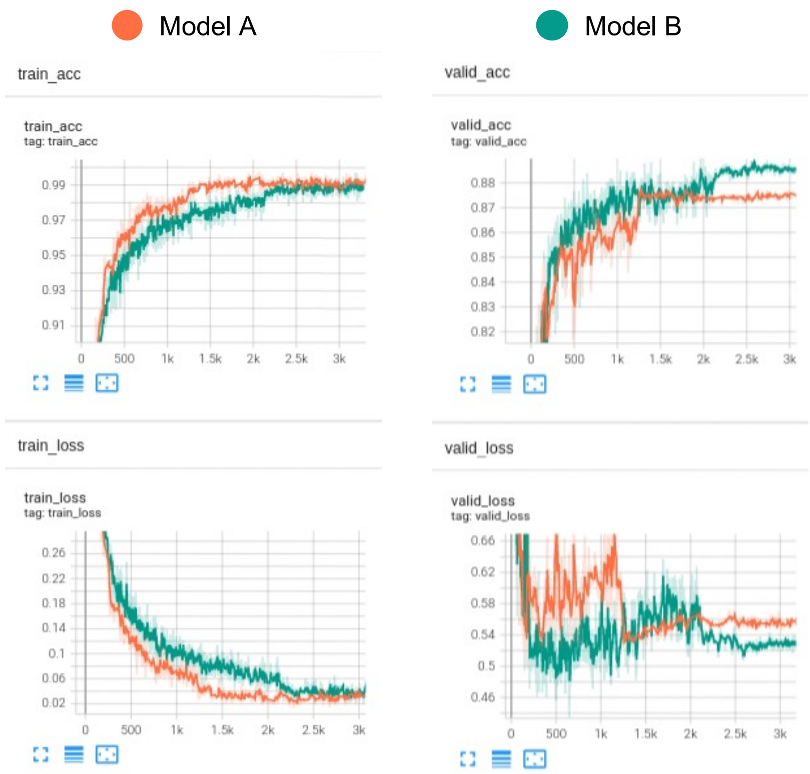
	(a)	(b)	(c)	(d)
训练集准确率	99.30%	99.70%	99.30%	99.40%
验证集准确率	88.15%	88.98%	88.63%	89.07%

总结学习曲线以及准确率对比可见，(a)(b)在学习率调整后均有约 1%的准确率提升，说明收敛到了更优的局部极值；(a)收敛速度、最终准确率均较差，这也与步长参数的选取有关；(b)由于在曲线平坦时再下降学习率，可达到较高的准确率，收敛速度与(a)以及恒定学习率时一致；(c)(d)使用余弦退火，由于学习率动态调整，使得网络能够较快找到更优的极值位置，尤其是验证集上准确率的收敛

速度较快，也能达到较好效果，其中(d)策略可达到最好的准确率，但由于本次任务较为简单，(d)达到较高准确率所用 epoch 略多于(b)。综合考虑，选择(b)作为较优的学习率策略。

(3) 结合数据增强和学习率策略，提升网络性能。

综合(1)(2)的实验，最终使用(i)+(iv)的数据增强和(b)的学习率策略，训练网络至 3000 个 epoch，得到模型 B 在训练集最高准确率为 **99.30%**，验证集最高准确率为 **88.96%**。得到训练、验证曲线并与相同条件下训练的模型 A 对比如下。



由此可见，与模型 A 相比，最终的验证集准确率较高，同时学习曲线较为平滑。模型 B 在训练初期的训练集准确率不及模型 A，但验证集准确率比模型 A 较好，这主要是因为模型 B 通道数、参数量更少，更不容易过拟合，而多 Pathway 等设计使得模型也能较好地学习到特征，并在验证集上有较好表现。

附：

附录文件夹包含了本次作业实现的全部代码。其中，CNN 实现中 main.py 第 183～186 行、第 143～144 行是不同学习率策略调整方法，可根据需要取消注释进行选择；data.py 第 18～24 行包含了不同数据增强方法，可根据需要取消注释

进行选择。

参考文献：

[1] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. CVPR, 2017.

[2] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In CVPR, 2022.