

基于热重启机制的胶囊投影网络快速训练算法

谢 奔¹, 张索非², 吴晓富¹

(1. 南京邮电大学 通信与信息工程学院, 江苏 南京 210003;

2. 南京邮电大学 物联网学院, 江苏 南京 210003)

摘 要: 胶囊投影网络是一种新型的深度神经网络结构, 将传统的卷积神经网络与胶囊投影结构结合起来学习潜在的视觉特征。尽管胶囊投影网络在多个分类数据集上展现出了先进的性能, 但是训练该算法模型往往需要较高的学习成本, 这对胶囊投影网络在实际问题中的应用带来一定的限制。针对该问题, 将基于热重启机制的随机梯度下降算法引入到胶囊投影网络的学习中, 提出了一种基于热重启机制的胶囊投影网络快速训练算法, 并在多个分类数据集上对该方法进行实验评估。实验结果表明, 与原始的胶囊投影网络相比, 该方法不仅解决了训练成本高昂的问题, 同时所学模型也具有比较好的泛化性能。

关键词: 卷积神经网络; 胶囊投影结构; 热重启机制; 快速训练算法; 深度学习

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2020)12-0021-06

doi: 10.3969/j.issn.1673-629X.2020.12.004

Learning Capsule Projection Network by Stochastic Gradient Descent with Warm Restarts

XIE Ben¹, ZHANG Suo-fei², WU Xiao-fu¹

(1. School of Telecommunications and Information Engineering, Nanjing University of Posts and

Telecommunications, Nanjing 210003, China;

2. School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Capsule projection network (CapProNet) is a recently proposed deep neural network architecture, which provides potential features by combining conventional deep networks with capsule projection structure. Although CapProNet shows competitive performance on various benchmark datasets, the model requires much expensive budget for training, which brings certain limitations to the application of CapProNet in practical problems. To address this problem, we introduce stochastic gradient descent with warm restarts (SGDR) into the learning of CapProNet and propose a CapProNet model fast training algorithm based on the warm restarts. Different learning strategies of methods are compared and evaluated. The experiment demonstrates that the proposed method can deliver better generalization performance with equivalent or even less training epochs compared with the traditional training method.

Key words: convolutional neural network; capsule projection structure; warm restarts; fast training algorithm; deep learning

0 引言

在过去的十年, 卷积神经网络 (convolutional neural network, CNN) 在包括图像分类、语义分割、目标检测、人脸识别等计算机视觉任务上都取得了不错的成绩^[1-4]。尽管卷积神经网络在很多领域中都取得了重大突破, 但是它仍然存在着一定的局限性, 比如网络中的池化层会使得图像中的空间信息丢失等。为了解决上述问题, Hinton 等人提出了胶囊网络^[5], 并在几个小型分类数据集 (如 MNIST^[6]) 上取得了不错的性

能, 后续有一些工作者对其结构进行了研究^[7-9], 发现胶囊网络结构过于简单以及动态路由机制的计算过于复杂是导致其在大型分类数据集 (如 CIFAR10, CIFAR100^[10]) 上没有传统的卷积神经网络性能好的原因。针对胶囊网络的这些局限性, 文献[11]提出一种胶囊投影网络 (capsule projection network, CapProNet)。该模型使用一组胶囊张成的子空间来表征实体, 一旦一组胶囊子空间被学习到, 模型即可将提取的特征投影到这组胶囊子空间上形成一组胶囊以用

收稿日期: 2020-01-10

修回日期: 2020-05-13

基金项目: 国家自然科学基金 (61372123, 61701252)

作者简介: 谢 奔 (1996-), 男, 硕士研究生, 通信作者, 研究方向为深度学习与计算机视觉; 张索非, 博士, 讲师, 研究方向为图像与视频信号处理、机器学习、物联网技术等; 吴晓富, 博士, 教授, 研究方向为信息论与编码、机器学习与计算机视觉、密码学与信息安全。

于分类任务。与文献[5]一样,胶囊的长度用于计算输入样本属于对应类别的概率,其方向则用于表示实体的一些其他属性,例如姿势、尺寸、纹理等。胶囊投影网络的另一个优点在于其可以很容易地嵌入到现有的神经网络结构上,融入了胶囊投影网络后的结构在多个分类数据集上取得了性能提升。

尽管胶囊投影网络在多个分类数据集上取得了不错的性能,但是这是建立在比较高的训练成本基础上达到的,而过高的训练成本会削弱模型的泛化能力。为了解决上述问题,受文献[12]的启发,该文引入一种基于热重启机制的随机梯度下降算法(stochastic gradient descent with warm restarts, SGDR)到胶囊投影网络的学习中,提出了一种基于热重启机制的胶囊投影网络快速训练算法,并在多个分类数据集上进行评估。

实验结果表明,该方法在降低训练成本的同时实现了比较好的泛化性能,例如,使用 WideResNet^[13] 作为主干网与 CapProNet 结合后的网络模型在 CIFAR10 与 CIFAR100 数据集上分别取得的最好结果为 3.56% 与 18.73%, 然而文献[11]中报道的在 CIFAR10 与 CIFAR100 数据集上分别取得的最好结果只有 3.64% 与 19.83%。对应地,该方法使得模型训练完全收敛只需要 300 个 epochs, 相较于原论文的训练策略所需

要的 500 个 epochs, 很大程度上降低了训练成本。

1 胶囊投影网络

1.1 网络结构

胶囊投影网络是文献[11]中提出的一种新型神经网络结构, 不同于文献[5, 14], 胶囊投影网络通过学习一组胶囊子空间表征一组实体类, 胶囊子空间的学习是为了将主干网提取的特征投影到上面形成一组胶囊用于分类任务。同样地, 胶囊的长度可以用于计算输入样本属于对应类别的概率, 其方向则用于表示实体的一些其他属性, 例如姿势、尺寸、纹理等。与原始胶囊网络结构不同的是, 胶囊投影网络可以作为分类层(classification layer)嵌入到多种现有的神经网络结构(例如 ResNet^[15]、WideResNet^[13] 等)中来帮助网络更好地学习特征, 以提高网络的性能。

一个典型的胶囊投影网络与主干网络相结合的结构如图1所示。这里以 CIFAR10 数据集为例, 首先, 给定输入图片, 经过一个主干网提取出特征 $x \in \mathbb{R}^d$; 然后, 当胶囊投影网络学习到 10 个胶囊子空间 $S = \{S_1, S_2, \dots, S_{10}\}$ 后, 可以将特征 x 分别投影到这些子空间上得到 10 个胶囊 $[v_1, v_2, \dots, v_{10}] \in \mathbb{R}^{10 \times d}$, 这里 d 指胶囊的维度;

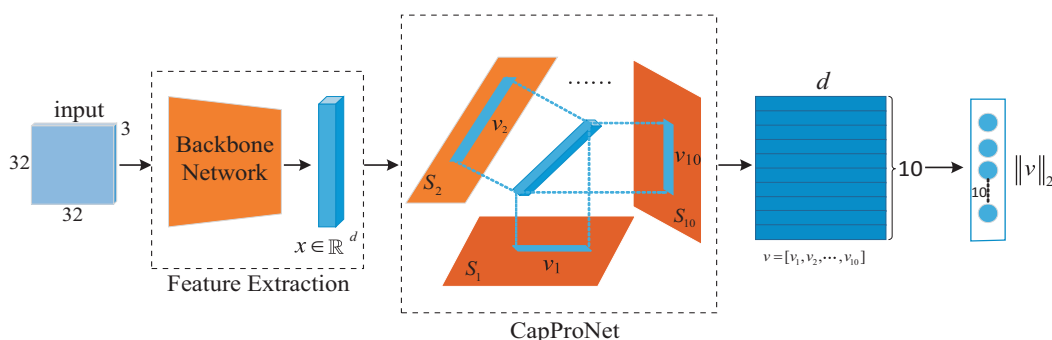


图1 主干网络与胶囊投影网络结合后的网络结构

最后, 可以通过计算这些胶囊的长度 $\|v\|_2$ 得到输入图片属于对应类别的概率, 其中长度最长的胶囊对应最终的分类结果。

1.2 特征在胶囊子空间上的投影

处于网络结构核心位置的是胶囊投影算法, 这一部分将详细对该算法进行一个说明。假设输入样本有 L 类, 这样胶囊投影网络需要学习 L 个胶囊子空间, 即 $\{S_1, S_2, \dots, S_L\}$ 。假设主干网提取出输入样本的特征 $x \in \mathbb{R}^d$, 为了学习到一定的特征表示, 将 x 分别投影到 L 个胶囊子空间上, 从而得到 L 个胶囊 $[v_1, v_2, \dots, v_L]$, 每个胶囊的长度用于表示输入样本属于对应类别的概率。

对于每个胶囊子空间而言, 网络通过学习一个权

重矩阵 $W_i = \{w_1, w_2, \dots, w_c\} \in \mathbb{R}^{d \times c}$, 使用权重矩阵的列向量作为基向量构建子空间, 即 $S_i = \text{span}(W_i)$ 。这样特征在胶囊子空间上的投影可通过式(1)得到:

$$\begin{cases} v_i = P_i x \\ P_i = W_i W_i^+ \end{cases} \quad (1)$$

其中, P_i 指投影矩阵, W_i^+ 指广义逆矩阵^[16]。当 W_i^+ 的列向量相互独立时, W_i^+ 可以近似为 $(W_i^T W_i)^{-1} W_i^T$, 这样可以通过式(2)计算胶囊的长度 $\|v_i\|_2$:

$$\|v_i\|_2 = \sqrt{v_i^T v_i} = \sqrt{x^T P_i^T P_i x} = \sqrt{x^T W_i \Sigma_i W_i^T x} \quad (2)$$

其中, $\Sigma_i = (W_i^T W_i)^{-1}$ 指归一化矩阵, 在实际网络进行参数更新时, 该矩阵对于网络学习新的特征信息有很大帮助。

2 基于热重启机制的胶囊投影网络快速训练算法

基于热重启机制的梯度下降算法 (SGDR)^[12] 通过将 SGD 优化器与热重启机制相结合在多个分类任务上取得了不错的效果。所谓的热重启机制是指在网络训练的时候人为设定一个重启周期 T_i , 每达到这个周期, 将重新启动初始学习率进行训练, 同时模型的每次重启使用上一次重启的参数作为初始化, 其本质上就是一个学习率的调整。该文通过将 SGDR 算法与胶囊投影网络相结合, 提出了一种基于热重启机制的胶囊投影网络快速训练算法。

对于基于热重启机制的胶囊投影网络快速训练算法, 采用余弦退火方案来调整学习率, 从而达到热重启的作用。对于每次迭代, 采用式 (3) 调整学习率:

$$\eta_t = \eta_{\min}^{(i)} + \frac{1}{2}(\eta_{\max}^{(i)} - \eta_{\min}^{(i)})(1 + \cos(\frac{T_{\text{cur}}}{T_i}\pi)) \quad (3)$$

其中, i 表示迭代次数, $\eta_{\max}^{(i)}$ 与 $\eta_{\min}^{(i)}$ 分别表示学习率调整的上下界, T_{cur} 表示每次重启开始到结束这一区间 epoch 的变化次数, T_i 表示重启周期。从式 (3) 可以看出, 当 $t=0$ 时, $\eta_t = \eta_{\max}^{(i)}$; 当 $T_{\text{cur}} = T_i$ 时, 余弦函数输出为 -1 , 因此此时的 $\eta_t = \eta_{\min}^{(i)}$ 。在每次重启时通过调整 $\eta_{\max}^{(i)}$ 与 $\eta_{\min}^{(i)}$ 可能带来胶囊投影网络性能的提高, 但是这会引入额外的超参数, 所以不考虑这个选择。文中令 $\eta_{\max}^{(i)} = 1$, $\eta_{\min}^{(i)} = 0$, 这样式 (3) 可以简化为如下形式:

$$\eta_t = \frac{1}{2}(1 + \cos(\frac{T_{\text{cur}}}{T_i}\pi)) \quad (4)$$

为了使基于热重启机制的胶囊投影网络快速训练算法更好地提升模型性能, 初始设置 $T_{i=0}$ 为一个相对较小的值, 例如将其设置为总的训练预算的 $1/10$, 在下一次重启前再乘上一个乘数因子 T_{mult} , 以变化重启周期 T_i 。通过这种动态改变重启周期的方式来促使网络模型快速找到最优解, 以降低训练成本。

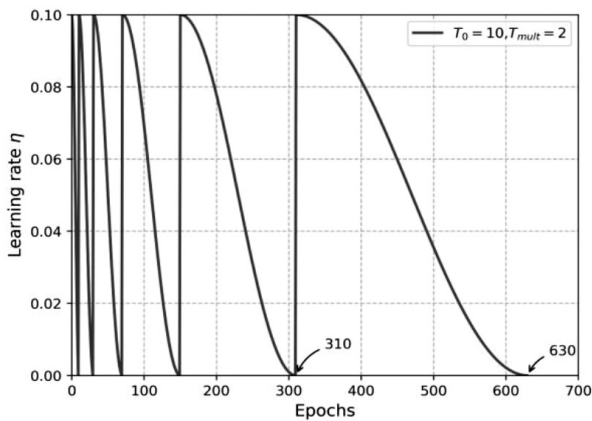


图 2 算法示例

这里, 通过一个例子来说明基于热重启机制的胶

囊投影网络快速训练算法的具体设置, 如图 2 所示。将初始学习率设为 0.1 , 同时令 $T_{i=0} = 10$, $T_{\text{mult}} = 2$, 当训练的 epoch 达到 10 时, 此时学习率 $\eta_{t=10} = 0$, 这时由于 $T_{\text{cur}} = T_{i=0}$, 所以在下一次重启前需要将 T_{cur} 设置为 0 , 这样学习率 η_t 将重新回到初始学习率, 网络也将重新回到初始状态进行训练, 同时将 $T_{i=0}$ 乘上一个乘数因子 $T_{\text{mult}} = 2$, 从而得到下一次重启的周期为 $T_{i=1} = T_{i=0} T_{\text{mult}} = 20$, 以此类推, 直到训练结束。基于热重启机制的胶囊投影网络快速训练算法通过这种反复重启来促使网络模型快速找到最优解, 从而在提高网络的性能的同时也降低了训练成本。

3 实验

本节通过实验来验证基于热重启机制的胶囊投影网络快速训练算法可以有效地提升网络模型性能。

3.1 数据集

实验中采用 CIFAR 数据集来评估所提出的方法。CIFAR 数据集包含 $50\,000$ 张像素为 32×32 的图片作为训练集, $10\,000$ 张像素为 32×32 的图片作为测试集。CIFAR10 数据集集中的所有图片来源于 10 类目标, CIFAR100 数据集集中的所有图片来源于 100 类目标。

3.2 主干网络

实验中选取多种主干网络来评估性能, 例如 ResNet^[15]、WideResNet^[13]、Densenet^[17] 和 ResNet (pre-activation)^[18]。将主干网络的最后一层分类层用胶囊投影网络 (CapProNet) 替代, 同时将结合后的网络在所给数据集上进行端到端训练, 训练后的模型在测试集上测试得出测试结果。为了书写方便, 将主干网标记为 X , 这样所有主干网络与胶囊投影网络结合后的网络结构可以记为 CapProNet+ X 。

3.3 不同 SGDR 配置对性能的影响分析

根据第 2 节的介绍可以发现, 不同的 SGDR 设置对网络模型性能影响很大, 同时由于 SGDR 设置的不同, 模型在收敛的时候所需要的 epochs 也不一样。例如, 当 $T_{i=0} = 10$, $T_{\text{mult}} = 2$ 时, 根据图 2 可以看出, 模型可能是在第 310 个 epoch 收敛的比较好, 又比如, 当 $T_{i=0} = 50$, $T_{\text{mult}} = 1$ 时, 模型可能是第 400 个 epoch 收敛的比较好, 所以为了找到促使胶囊投影网络收敛最好, 同时训练成本也较少的设置, 就不同 SGDR 配置对模型性能的影响做了一个消融研究。

选取以下几种设置进行消融研究: (1) $T_{i=0} = 1$, $T_{\text{mult}} = 2$; (2) $T_{i=0} = 10$, $T_{\text{mult}} = 2$; (3) $T_{i=0} = 20$, $T_{\text{mult}} = 2$; (4) $T_{i=0} = 50$, $T_{\text{mult}} = 1$; (5) $T_{i=0} = 100$, $T_{\text{mult}} = 1$; (6) $T_{i=0} = 200$, $T_{\text{mult}} = 1$, 使用 CapProNet ($c = 2$) + ResNet110 网络结构在 CIFAR10 数据集上进行训练测试, 测试结果如表 1 所示。

表 1 不同 SGDR 设置条件下 CapProNet($c=2$) + ResNet110 网络在 CIFAR10 上的测试错误率

Settings	Params	epochs	CIFAR10
$T_{i=0} = 1, T_{mult} = 2$	1.7 M	511	5.20
$T_{i=0} = 10, T_{mult} = 2$	1.7 M	310	5.05
$T_{i=0} = 20, T_{mult} = 2$	1.7 M	300	5.43
$T_{i=0} = 50, T_{mult} = 1$	1.7 M	400	5.14
$T_{i=0} = 100, T_{mult} = 1$	1.7 M	400	5.10
$T_{i=0} = 200, T_{mult} = 1$	1.7 M	400	5.07

从表 1 可以发现, 尽管 $T_{i=0} = 20, T_{mult} = 2$ 设置的训练成本最低, 但是其在 CIFAR10 数据集上的分类错误率是这几种设置中最高的, 相比之下, $T_{i=0} = 10, T_{mult} = 2$ 设置在比其训练成本多 10 个 epochs 的条件下在 CIFAR10 数据集上获得了最低的错误率, 错误率达到了 5.05%, 这样在降低训练成本的同时获得了比较好的性能。所以在接下来的所有实验中采用了 $T_{i=0} = 10, T_{mult} = 2$ 设置来训练网络模型。

3.4 CIFAR100 数据集上的分类结果

评估了 CapProNet+WideResNet28-10、CapProNet+ResNet110、CapProNet+Densenet100-12 和 CapProNet+ResNet164(pre-activation) 四种网络结构在 CIFAR100 上的性能。实验结果如表 2 所示, 与原论文报道相比, 该方法实现了更高的性能, 在 CIFAR100 上最好的性能达到 3.56%, 相比于原论文, 减少了 2% 的错误率。更重要的是比原训练框架减少了接近 40% 的训练成本, 原训练框架需要 500 个 epochs, 而我们的仅需要大约 300 个 epochs, 这在很大程度上降低了训练成本。

表 2 不同网络结构在 CIFAR100 上的测试错误率

Method	epochs	c	CIFAR100
CapProNet+ResNet110			
original	500	2	5.24
ours	310	2	5.05
original	500	4	5.27
ours	310	4	5.02
original	500	8	5.19
ours	310	8	5.08
CapProNet+WideResNet28-10			
original	500	4	3.64
ours	310	4	3.56
original	500	8	3.84
ours	310	8	3.83
CapProNet+ResNet164 (pre-activation) on			
original/rerun	500	4	4.88/5.18
ours	310	4	5.17
original/rerun	500	8	4.89/5.09
ours	310	8	5.11

续表 2

Method	epochs	c	CIFAR100
CapProNet+Densenet100-12			
original/rerun	500	4	4.35/4.64
ours	310	4	4.70
original/rerun	500	8	4.25/4.87
ours	310	8	4.88

注意对于 CapProNet + Densenet100-12 和 CapProNet+ResNet164(pre-activation) 网络结构, 用 500 个 epoch 无法复现原论文报道的结果, 所以为了保证公平比较, 在表 2 中报道了原论文报道结果与复现的结果, 标记为 original/rerun, 尽管所提方法并没有达到比较好的性能, 但是用更少的训练成本达到了与原论文相当的结果。

3.5 CIFAR100 数据集上的分类结果

同样评估了 CapProNet+WideResNet28-10、CapProNet+ResNet110、CapProNet+Densenet100-12 三种网络结构在 CIFAR100 上的性能。实验结果如表 3 所示, 与 3.4 节一样, 也报道了原论文与复现的结果, 标记为 original/rerun。从表中可以看到, 在 CIFAR100 数据集上最好的性能是 18.73%, 相比于原论文, 降低了 6% 的错误率, 更重要是, 该方法比原训练框架减少了接近 40% 的训练成本, 在减少训练的总预算的同时, 提升了模型的性能。

表 3 不同网络结构在 CIFAR100 上的测试错误率

Method	epochs	c	CIFAR100
CapProNet+ResNet110			
original/rerun	500	2	22.65/25.98
ours	310	2	25.70
original/rerun	500	4	22.45/26.94
ours	310	4	25.51
original/rerun	500	8	21.93/26.74
ours	310	8	25.02
CapProNet+WideResNet28-10			
original	500	4	19.98
ours	310	4	18.73
original	500	8	19.83
ours	310	8	18.92
CapProNet+Densenet100-12			
original	500	4	21.22
ours	310	4	21.08
original	500	8	21.19
ours	310	8	21.17

3.6 学习曲线的可视化

为了更好地展现提出的算法的有效性,将使用提出的算法的网络模型的学习曲线与原论文网络模型进行一个对比。这里选取了 CapProNet+ResNet164 ($c=4$) 与 CapProNet+Densenet ($c=4$) 在 CIFAR10 数据集上训练得到的模型。

图 3 与图 4 分别为 CapProNet+ResNet164 ($c=4$) 与 CapProNet+Densenet ($c=4$) 在 CIFAR10 数据集上的学习曲线,其中 ours 表示使用提出的算法学习得到的模型,original 则表示原论文的模型。由图可见,使用提出的算法的模型可以更快地寻找到最优解,相比于原论文的模型需要 500 个 epochs 才能完全收敛,而提出的算法第 310 个 epochs 就可以完全收敛,这也进一步证明了提出的算法的有效性。

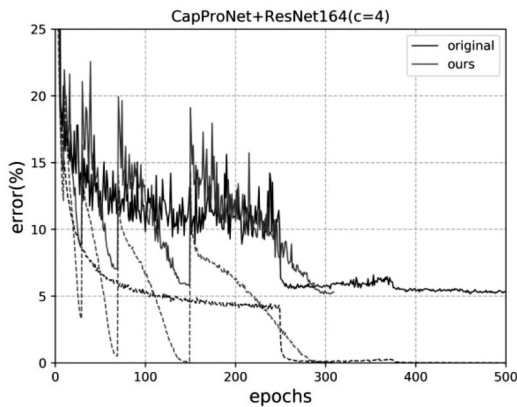


图 3 CapProNet+ResNet164 ($c=4$) 模型的学习曲线

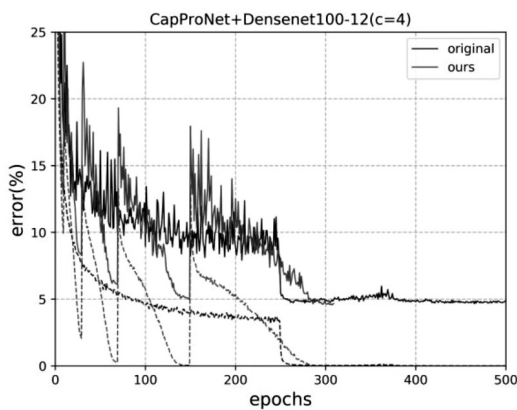


图 4 CapProNet+Densenet100-12 ($c=4$) 模型的学习曲线

3.7 消融研究

在这一部分,进一步探究在网络训练过程中通过增加 epochs 来获得更高性能的方法。考虑以下三种网络结构: CapProNet+WideResNet28-10、CapProNet+ResNet110、CapProNet+Densenet100-12,将所有的模型在 CIFAR10 数据集上使用 630 个 epochs 训练,同时为了公平比较,使用相同的训练成本重新训练原论文中的模型,二者实验的超参数保持一致。实验结果如表

4 所示,总体来说,提出的算法实现了比原论文更好的性能,在 CIFAR10 上最好的性能达到了 3.41%,相比文献[4]报道的最好成绩降低了 6% 的错误率。这些结果表明,提出的方法使用更高的训练成本可以达到更好的泛化性能。

表 4 不同网络结构在 CIFAR10 上的测试错误率

Method	epochs	c	CIFAR10
CapProNet+ResNet110			
rerun	630	2	5.09
ours	630	2	4.87
rerun	630	4	5.27
ours	630	4	5.07
rerun	630	8	5.03
ours	630	8	4.88
CapProNet+WideResNet28-10			
rerun	630	4	3.64
ours	630	4	3.41
rerun	630	8	3.76
ours	630	8	3.83
CapProNet+Densenet100-12			
rerun	630	4	4.64
ours	630	4	4.59
rerun	630	8	4.71
ours	630	8	4.67

3.8 SGDR 算法的兼容性

实验证明,所提出的算法在收敛速度与分类性能上均超过了原胶囊投影网络模型。但是这并不能代表 SGDR 算法应用于所有深度神经网络结构都可以收敛的很好,为了证明算法不具有普遍应用性,将 SGDR 算法应用于 ResNet110 与 ResNet164 (pre-activation) 两种结构上,不同结构在 CIFAR10 上测试的结果如表 5 所示。正如表 5 所示,相同的网络模型下,使用 SGDR 算法的结果远不如不用 SGDR 算法的结果,这也证明了 SGDR 算法并不适用于所有深度神经网络结构。

表 5 不同训练策略在 CIFAR10 上的测试错误率

Method	epochs	SGDR	CIFAR10
ResNet110	310	No	5.43
ResNet110	310	Yes	6.55
ResNet164 (pre-activation)	310	No	5.60
ResNet164 (pre-activation)	310	Yes	5.72

4 结束语

将基于热重启机制的随机梯度下降算法引入到胶囊投影网络的学习中,提出了一种基于热重启机制的

胶囊投影网络快速训练算法,并分别在 CIFAR10 与 CIFAR100 数据集上进行了性能评估。实验结果表明,该方法在降低训练成本的同时还获得了比较好的泛化性能,因此,基于热重启机制的随机梯度下降算法可作为胶囊投影网络的一种标准训练算法,推动这种网络模型的进一步发展以及应用。目前的胶囊投影网络只用于分类任务中,在未来可以进一步探究其在目标检测、人脸识别与行人重识别等任务的可能性。

参考文献:

- [1] 杨真真,匡楠,范露,等. 基于卷积神经网络的图像分类算法综述[J]. 信号处理,2018,34(12):1474-1489.
- [2] 王嫣然,陈清亮,吴俊君. 面向复杂环境的图像语义分割方法综述[J]. 计算机科学,2019,46(9):36-46.
- [3] 许必霄,宫婧,孙知信. 基于卷积神经网络的目标检测模型综述[J]. 计算机技术与发展,2019,29(12):87-92.
- [4] 陆兴华,王凌丰,曾世豪,等. 基于神经网络学习的多姿态人脸图像识别算法[J]. 计算机技术与发展,2019,29(11):57-61.
- [5] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[C]//Advances in neural information processing systems. Long Beach, CA, USA:[s. n.], 2017:3856-3866.
- [6] DENG L. The mnist database of handwritten digit images for machine learning research [best of the web][J]. IEEE Signal Processing Magazine, 2012, 29(6):141-142.
- [7] 范文豪,吴晓富,张索非. 动态路由胶囊网络的可视化研究[J]. 计算机技术与发展,2019,29(8):71-75.
- [8] 付家慧,吴晓富,张索非. 基于仿射变换的胶囊网络特征研究[J]. 信号处理,2018,34(12):1508-1516.
- [9] 张天柱,邹承明. 使用模糊聚类的胶囊网络在图像分类上的研究[J]. 计算机科学,2019,46(12):279-285.
- [10] KRIZHEVSKY A, NAIR V, HINTON G. Cifar-10 (canadian institute for advanced research) [EB/OL]. 2010. <http://www.cs.toronto.edu/kriz/cifar.html>.
- [11] ZHANG L, EDRAKI M, QI G J. CapProNet: deep feature learning via orthogonal projections onto capsule subspaces [C]//Advances in neural information processing systems. Long Beach, CA, USA:[s. n.], 2018:5814-5823.
- [12] DONTI P, AMOS B, KOLTER J Z. Task-based end-to-end model learning in stochastic optimization[C]//Advances in neural information processing systems. Long Beach, CA, USA:[s. n.], 2017:5484-5494.
- [13] ZERHOUNI E, LÁNYI D, VIANA M, et al. Wide residual networks for mitosis detection[C]//2017 IEEE 14th international symposium on biomedical imaging. [s. l.]: IEEE, 2017:924-928.
- [14] HINTON G E, KRIZHEVSKY A, WANG S D. Transforming auto-encoders [C]//International conference on artificial neural networks. Berlin:Springer, 2011:44-51.
- [15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2016:770-778.
- [16] BEN-ISRAEL A, GREVILLE T N E. Generalized inverses: theory and applications [M]. [s. l.]: Springer Science & Business Media, 2003.
- [17] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2017:4700-4708.
- [18] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks [C]//European conference on computer vision. Cham:Springer, 2016:630-645.