Born-Again

2018年8月14日 15:00

参考： https://danieltakeshi.github.io/2018/05/27/bann/

通过ERM(Empricial Risk Minimization)产生result model θ_1^*，最小化loss

loss function:

$$\theta_1^* = \arg\min_{\theta_1} \mathcal{L}(y, f(x, \theta_1)), \qquad (1)$$

这篇论文认为

$$f(x, \theta_1^*)$$

包含了丰富的信号。新的loss function表示为

$$\mathcal{L}(f(x, \arg\min_{\theta_1} \mathcal{L}(y, f(x, \theta_1))), f(x, \theta_2)). \qquad (2)$$

(就是新模型的输出和旧模型输出的交叉熵)

3.1. Sequence of Teaching Selves Born-Again Networks Ensemble

第k个模型的loss function。

$$\mathcal{L}(f(x, \arg\min_{\theta_{k-1}} \mathcal{L}(f(x, \theta_{k-1}))), f(x, \theta_k)). \qquad (3)$$

产生一个Born-Again Network Ensembles（BANE）

$$\hat{f}^k(x) = \sum_{i=1}^{k} f(x, \theta_i)/k. \qquad (4)$$

序列学习的提升饱和后，ensemble可以有显著提升。

3.2. Dark Knowledge Under the Light

Dark Knowledge：隐含在wrong response里面的分布，正如Hinton论文指出的，隐含了类别的相似性信息。

z：student logits，**t**：teacher logits

x：input samples

$Z = \sum_{k=1}^{n} e^{z_k}$ and $T = \sum_{k=1}^{n} e^{t_k}$

交叉熵：

$$\mathcal{L}(\mathbf{x}_1, \mathbf{t}_1) = -\sum_{k=1}^{n} \left(\frac{e^{t_k}}{T} \log \frac{e^{z_k}}{Z} \right)$$

然后求偏导：

$$\frac{\partial \mathcal{L}_i}{\partial z_i} = q_i - p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} - \frac{e^{t_i}}{\sum_{j=1}^{n} e^{t_j}}. \qquad (5)$$

现在考虑有***b***个元素的minibatch，有：

$$\mathcal{L}(\mathbf{x}_1, \mathbf{t}_1, \ldots, \mathbf{x}_b, \mathbf{t}_b) = \frac{1}{b} \sum_{s=1}^{b} \mathcal{L}(\mathbf{x}_s, \mathbf{t}_s)$$

求偏导：（假设*n*th label就是ground truth label *）

$$\sum_{s=1}^{b} \sum_{i=1}^{n} \frac{\partial \mathcal{L}_{i,s}}{\partial z_{i,s}} = \underbrace{\boxed{\sum_{s=1}^{b} (q_{*,s} - p_{*,s})}}_{\text{gradient from correct choice}} + \underbrace{\boxed{\sum_{s=1}^{b} \sum_{i=1}^{n-1} (q_{i,s} - p_{i,s})}}_{\text{wrong output.}}, \qquad (7)$$

除以b，（7）里面的第一项重写成：

$$\frac{1}{b} \sum_{s=1}^{b} (q_{*,s} - p_{*,s} y_{*,s}) \qquad (8)$$

$$\frac{\partial \mathcal{L}_*}{\partial z_*} = q_* - y_* = \frac{e^{z_*}}{\sum_{j=1}^{n} e^{z_j}} - 1 \qquad (6)$$

teacher输出p_*就是ground truth labe y_*的一个权重

这个可以理解为给一个batch里的样本进行重要性加权。老师正确 & 对输出有confidence的时候，$p_* \approx 1$。就变成（6）了。老师有比较少的confidence，这个样本对整个训练的contribution变少。

于是提出importance weights w_s. When the importance weights correspond to the output of a teacher for the correct dimension we have:

$$\sum_{s=1}^{b} \frac{w_s}{\sum_{u=1}^{b} w_u} (q_{*,s} - y_{*,s}) = \sum_{s=1}^{b} \frac{p_{*,s}}{\sum_{u=1}^{b} p_{*,u}} (q_{*,s} - y_{*,s}). \quad (9)$$

提出疑问：dark knowledge的成功之处在于 1.老师的non-argmax输出（wrong output） 还是 2. 一种importance weighting?

$$\sum_{s=1}^{b} \sum_{i=1}^{n} \frac{\partial \mathcal{L}_{i,s}}{\partial z_{i,s}} = \sum_{s=1}^{b} (q_{*,s} - p_{*,s}) + \sum_{s=1}^{b} \sum_{i=1}^{n-1} (q_{i,s} - p_{i,s}), \qquad (7)$$

→ 好：importance-weighting贡献大

→ 不好：non-argmax的信息很重要

提出两种treatments：

1. Confidence Weighted by Teacher Max (CWTM)

$$\sum_{s=1}^{b} \frac{\max p_{.,s}}{\sum_{u=1}^{b} \max p_{.,u}} (q_{*,s} - y_{*,s}). \qquad (10)$$

$$\sum_{s=1}^{b} \frac{\overset{\text{max}}{p_{*,s}}}{\sum_{u=1}^{b} p_{*,u}} (q_{*,s} - y_{*,s}). \quad (9)$$

2. Permuted Predictions (DKPP) 与1类似.

we permute the non-argmax outputs of the teacher's predicted distribution(除了argmax随机重排)

$$\sum_{s=1}^{b} \sum_{i=1}^{n} \frac{\partial \mathcal{L}_{i,s}}{\partial z_{i,s}} = \sum_{s=1}^{b} (q_{*,s} - \max p_{.,s}) + \sum_{s=1}^{b} \sum_{i=1}^{n-1} q_{i,s} - \underbrace{\phi(p_{j,s})}_{\text{permuted output}}, \qquad (11)$$

$$\sum_{s=1}^{b} \sum_{i=1}^{n} \frac{\partial \mathcal{L}_{i,s}}{\partial z_{i,s}} = \sum_{s=1}^{b} (q_{*,s} - \overset{\text{max}}{p_{*,s}}) + \sum_{s=1}^{b} \sum_{i=1}^{n-1} (q_{i,s} - \overset{\text{permute}}{p_{i,s}}), \qquad (7)$$

实验

Table 1. **Test error on CIFAR-10** for Wide-ResNet with different depth and width and DenseNet of different depth and growth factor

Network	Parameters	Teacher	BAN
Wide-ResNet-28-1	0.38 M	6.69	**6.64**
Wide-ResNet-28-2	1.48 M	5.06	**4.86**
Wide-ResNet-28-5	9.16 M	4.13	**4.03**
Wide-ResNet-28-10	36 M	**3.77**	3.86
DenseNet-112-33	6.3 M	3.84	**3.61**
DenseNet-90-60	16.1 M	3.81	**3.5**
DenseNet-80-80	22.4 M	**3.48**	3.49
DenseNet-80-120	50.4 M	**3.37**	3.54

BAN

不同深度的架构

之间的差距迅速缩小.

construct comparable ResNet students by switching Dense Blocks with Wide Residual Blocks and Bottleneck Residual Blocks

Table 2. **Test error on CIFAR-100** *Left Side:* DenseNet of different depth and growth factor and respective BAN student. BAN models are trained only with the teacher loss, BAN+L with both label and teacher loss. CWTM are trained with sample importance weighted label, the importance of the sample is determined by the max of the teacher's output. DKPP are trained only from teacher outputs with all the dimensions but the argmax permuted. *Right Side:* test error on CIFAR-100 sequence of BAN-DenseNet, and the BAN-ensembles resulting from the sequence. Each BAN in the sequence is trained from cross-entropy with respect to the model at its left. BAN and BAN-1 models are trained from Teacher but have different random seeds. We include the teacher as a member of the ensemble for Ens*3 for 80-120 since we did not train a BAN-3 for this configuration.

depth growth factor　只有teacher　label　importance weight　permuted

Network	Teacher	BAN	BAN+L	CWTM	DKPP	BAN-1	BAN-2	BAN-3	Ens*2	Ens*3
DenseNet-112-33	18.25	**16.95**	17.68	17.84	17.84	17.61	17.22	**16.59**	15.77	15.68
DenseNet-90-60	17.69	**16.69**	16.93	17.42	17.43	16.62	**16.44**	16.72	15.39	15.74
DenseNet-80-80	17.16	**16.36**	16.5	17.16	16.84	16.26	16.30	**15.5**	15.46	15.14
DenseNet-80-120	16.87	~~**16.00**~~	16.41	~~17.12~~	16.34	**16.13**	16.13	/	**15.13**	**14.9**

不同random seed

SOTA

+ teacher

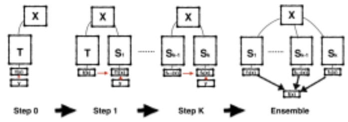
1.BAN中,小模型与大模型之间差距缩小.参数量18.

2.BAN-1到BAN-3.代与代之间的变化 EnS.

3.CWTM,DKPP与Teacher对比.

BAN概要

1. teacher和student使用相同的模型, student称为BAN, 结果表明BAN的表现明显优于teacher
2. 提出了两个distillation objectives: Confidence-Weighted by Teacher Max (**CWTM**) and (ii) Dark Knowledge with Permuted Predictions (**DKPP**). 探讨了teacher输出如何影响student的性能.
3. BAN细节:
    a. 训练流程:
        i. 训练teacher
        ii. teacher收敛以后初始化一个新的student, 训练它. student网络的训练目标是 1.预测正确label 2. 匹配teacher的输出分布
        iii. 训练好的student作为新的teacher, 训练下一代student, 如此重复
        iv. (可选)ensumble多个student



    b. 关于CWTM和DKPP
        i. Hinton认为蒸馏的成功可能是wrong responses携带了相似性信息导致的, 另外一个可能的解释是蒸馏可能类似于样本重要性加权, 其中权重对于teacher的confidence
        ii. 论文中有公式推导, 有点繁琐, 这里不太好说明. 最终推导的结果是student的梯度被分解成两项, 第一项是teacher的correct choice的梯度, 第二项是wrong outputs的梯度求和.
        iii. 作者认为老师的作用相当于给样本做了一个importance weights(第一项), 也可能是wrong outputs携带了一些类别之的相似性的信息(第二项)
        iv. 为了探讨到底是哪一项影响大, 提出了CWTM和DKPP, 都是对student的梯度做了改变.
    c. 实验结果
        i. 实验非常多, 有的不是特别懂. 数据集: CIFAR-10和CIFAR-100, 网络: Dense-Net(不同的超参: depth, growth, compression factors)
        ii. 老师和学生网络相同的实验结果:
            1) BAN的准确率高于Teacher
            2) 三代student, 整体上来看随着代数增加准确率上升(有部分例外情况)
            3) BAN中, 小的模型(参数少)和大模型的差距明显缩小
            4) CWTM和DKPP的结果显示, 蒸馏带来的性能提升主要来自于teacher对样本做了重要性加权
        iii. 还有其它学生和老师不同的实验, 因为对DenseNet不了解所以看得不是特别明白

有提升...不是单纯的靠 non-argmax knowledge

Table 3. **Test error on CIFAR-100** for Wide-ResNet students trained from identical Wide-ResNet teachers and for DenseNet-90-60 students trained from Wide-ResNet teachers

Network	Teacher	BAN	Dense-90-60
Wide-ResNet-28-1	30.05	29.43	24.93
Wide-ResNet-28-2	25.32	24.38	18.49
Wide-ResNet-28-5	20.88	20.93	17.52
Wide-ResNet-28-10	19.08	18.25	16.79

teacher: Wide-Res
students: Wide-Res

training a DenseNet-90-60 student from ResNet student confirms the trend of students surpassing their teachers.

Table 4. **Test error on CIFAR-100-Modified Densenet:** a Densenet-90-60 is used as teacher with students that share the same size of hidden states after each spatial transition but differs in depth and compression rate

Densenet-90-60	Teacher	0.5*Depth	2*Depth	3*Depth	4*Depth	0.5*Compr	0.75*Compr	1.5*compr
Error	17.69	16.95	16.43	16.64	16.64	19.83	17.3	18.89
Parameters	22.4 M	21.2 M	13.7 M	12.9 M	1 2.6 M	5.1 M	10.1 M	80.5 M

teacher: Dense-90-60
students

Table 5. **DenseNet to ResNet:** CIFAR-100 test error for BAN-ResNets trained from a DenseNet-90-60 teacher with different numbers of blocks and compression factors. In all the BAN architectures, the number of units per block is indicated first, followed by the ratio of input and output channels with respect to a DenseNet-90-60 block. All BAN architectures share the first (conv1) and last(fc-output) layer with the teacher which are frozen. Every dense block is effectively substituted by residual blocks

DenseNet 90-60	Parameters	Baseline	BAN
Pre-activation ResNet-1001	10.2 M	22.71	/
BAN-Pre-ResNet-14-0.5	7.3 M	20.28	18.8
BAN-Pre-ResNet-14-1	17.7 M	18.84	**17.39**
BAN-Wide-ResNet-1-1	20.9 M	20.4	19.12
BAN-Match-Wide-ResNet-2-1	43.1 M	18.83	17.42
BAN-Wide-ResNet-4-0.5	24.3 M	19.63	**17.13**
BAN-Wide-ResNet-4-1	87.3 M	18.77	17.18

Table 6. **Validation/Test perplexity on PTB** (lower is better) for BAN-LSTM language model of different complexity

Network	Parameters	Teacher Val	BAN+L Val	Teacher Test	BAN+L Test
ConvLSTM	19M	83.69	80.27	80.05	**76.97**
LSTM	52M	75.11	71.19	71.87	**68.56**