知識蒸留における Multiple Teacher Network の選択

ER17048 中島僚佑　　指導教授：藤吉弘亘

1. はじめに

知識蒸留は Teacher Network の知識を Student Network に伝えることで，精度低下を抑制しつつモデルサイズを圧縮する手法である．Teacher Network と Student Network のモデルサイズに大きなギャップがあると，Student Network の精度が向上しない問題がある．本研究では，サイズの異なる複数の Teacher Network を用いる Multiple Teacher Network を提案し，知識蒸留の高精度化を目的とする．

2. 知識蒸留と問題点

知識蒸留には，Teacher Network の出力を利用する Knowledge Distillation (KD)[1] と，中間情報を利用する Contrastive Representation Distillation(CRD)[2] がある．表 1 に Student Network を ResNet20 にした場合の KD による蒸留後の精度と Teacher Network のモデルサイズ (パラメータ数) を示す．Teacher Network が WRN40_2 より大きなモデルサイズになると，大幅な精度低下を招くことがわかる．本研究では，この問題が発生する Teacher Network を Large Teacher Network(LTN)，発生しない Teacher Network を Medium Teacher Network(MTN) と呼称する．この問題を解決する手法として，Teacher Network と Student Network 間に Teacher Assistant(TA) と呼ぶモデルを配置し，段階的に蒸留する手法 [3] が提案されているが，段階的な学習を必要とする．

表 1: ResNet20 に対する蒸留後精度

Teacher Network	WRN28_1	ResNet32	WRN28_2	WRN40_2	WRN28_6	WRN28_10
パラメータ数	371,347	466,457	1,147,251	2,248,991	13,155,603	36,497,171
精度	71.27	71.12	70.98	70.25	70.14	69.60

3. 提案手法

本研究では，図 1 に示すように複数の Teacher Network から蒸留する Multiple Teacher Network を提案する．このとき，どのように Teacher Network を選択すると効果的であるか調査する．提案手法では，Teacher Network の組み合わせを MTN+LTN とし，MTN に Student Network と LTN の橋渡し的な役割を期待する．また，MTN のみで蒸留を行った場合よりも，LTN を含めて蒸留することで，その知識も活用できると考える．

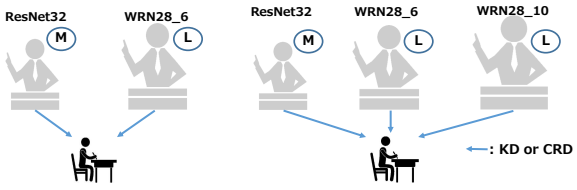

図 1: 提案手法の概念図

本研究での Student Network と Teacher Network 間の蒸留手法には，KD，CRD のいずれかを用いる．KD の損失関数を式 (1) に示す．

$$L_{KD} = \alpha L_{hard} + (1-\alpha)\tau^2 L_{soft} \tag{1}$$

ここで，α は係数，τ は温度パラメータを示す．CRD の損失関数を式 (2) に示す．

$$\begin{aligned} L_{CRD} &= \mathbb{E}_{q(g^T,g^S|C=1)}[\log \frac{e^{g^T g^S/\tau}}{e^{g^T g^S/\tau} + \frac{N}{M}}] \\ &+ N\mathbb{E}_{q(g^T,g^S|C=0)}[1 - \log\left(\frac{e^{g^T g^S/\tau}}{e^{g^T g^S/\tau} + \frac{N}{M}}\right)] \end{aligned} \tag{2}$$

ここで，τ は温度パラメータ，M はデータセットのカーディナリティで，データの種類がどのくらいあるかの度合いを表している．N は入力データの非類似ペアの数である．g^S と g^T は各ネットワークの出力層手前から抽出した出力を線形変換させ，L2 ノルムで正規化したものである．

また，最終的な損失は式 (3) に示す通り，全ての損失を総和した値になる．

$$L_{total} = \sum_{i=0}^{n} L_{iKDorCRD} \tag{3}$$

4. 評価実験

提案手法の有効性を評価するため，評価実験を行う．TA とも比較し評価する．

4.1. 実験概要

データセットには CIFAR-100 を使用し，学習時のエポック数は 200，バッチサイズは 64 とする．学習モデルは Student Network に ResNet20，LTN に WRN40_2〜WRN28_10，MTN に WRN28_1〜WRN28_2 を用いる．

4.2. 実験結果

表 2 は 2 つの Teacher Network を用いて蒸留した精度を示す．TN1，2，3 にはそれぞれ異なるモデルを使用し，TN3 は TN2 や TN1 より大きく，TN2 は TN1 より大きいモデルである．Teacher Network の組み合わせは，MTN のみ，MTN+LTN，LTN のみの 3 種類ある．

表 2: 2 つの Teacher Network で蒸留した精度

Teacher Network				提案手法	
	TN1	TN2	TA	KD	CRD
MTN のみ	ResNet32	WRN28_2	71.12	71.26	**72.13**
MTN+LTN	ResNet32	WRN28_6	70.83	71.24	71.73
MTN+LTN	ResNet32	WRN28_10	70.24	71.37	71.61
LTN のみ	WRN28_6	WRN28_10	69.37	69.92	70.75

表 2 から，MTN のみの場合が最も精度が高く，LTN のみの場合が最も低くなった．MTN+LTN の場合は，LTN のみより精度が高い．以上より，2 つの Teacher Network のうち 1 つを MTN にすることで，Student Network と LTN 間の蒸留で生じる問題を回避することができる．

表 3: 3 つの Teacher Network で蒸留した精度

Teacher Network					提案手法	
	TN1	TN2	TN3	TA	KD	CRD
MTN のみ	WRN28_1	ResNet32	WRN28_2	71.18	71.92	**72.19**
MTN+LTN	ResNet32	WRN28_2	WRN28_6	70.24	71.06	71.83
MTN+LTN	WRN28_2	WRN28_6	WRN28_10	68.87	70.15	71.00
LTN のみ	WRN40_2	WRN28_6	WRN28_10	68.99	70.01	70.55

表 3 は 3 つの Teacher Network を用いて蒸留した精度を示す．表 3 から，MTN+LTN の場合，精度低下を抑制することができないことがことがわかる．MTN のみの場合が最も高精度である．以上より，Teacher Network 数を増やしても LTN がある場合，高精度化が難しいことが判明した．

5. おわりに

本研究では，能力差のある Student Network と Teacher Network 間の蒸留の際に，MTN+LTN の組合せを含めてそれらを同時に蒸留させることを提案した．これにより，Teacher Network を 2 つ使用する場合は精度の低下を抑え，MTN のみの精度に匹敵し，さらに Teacher Network が 1 つのみで蒸留を行った場合よりも精度が向上することを確認した．また，TA と異なり段階的な学習を必要とせず，その分学習時間を短縮できることも確認した．

参考文献

[1] G. Hinton, *et al.*, “Distilling the knowledge in a neural network”, NeurIPS Workshop, 2015.

[2] Y. Tian, *et al.*, “Contrastive Representation Distillation”, ICLR, 2020.

[3] S. Mirzadeh, *et al.*, “Improved Knowledge Distillation via Teacher Assistant”, AAAI, 2020.