Born-Again Neural Networks

Ridge-i 論文よみかい

2018.07.26

Masanari Kimura

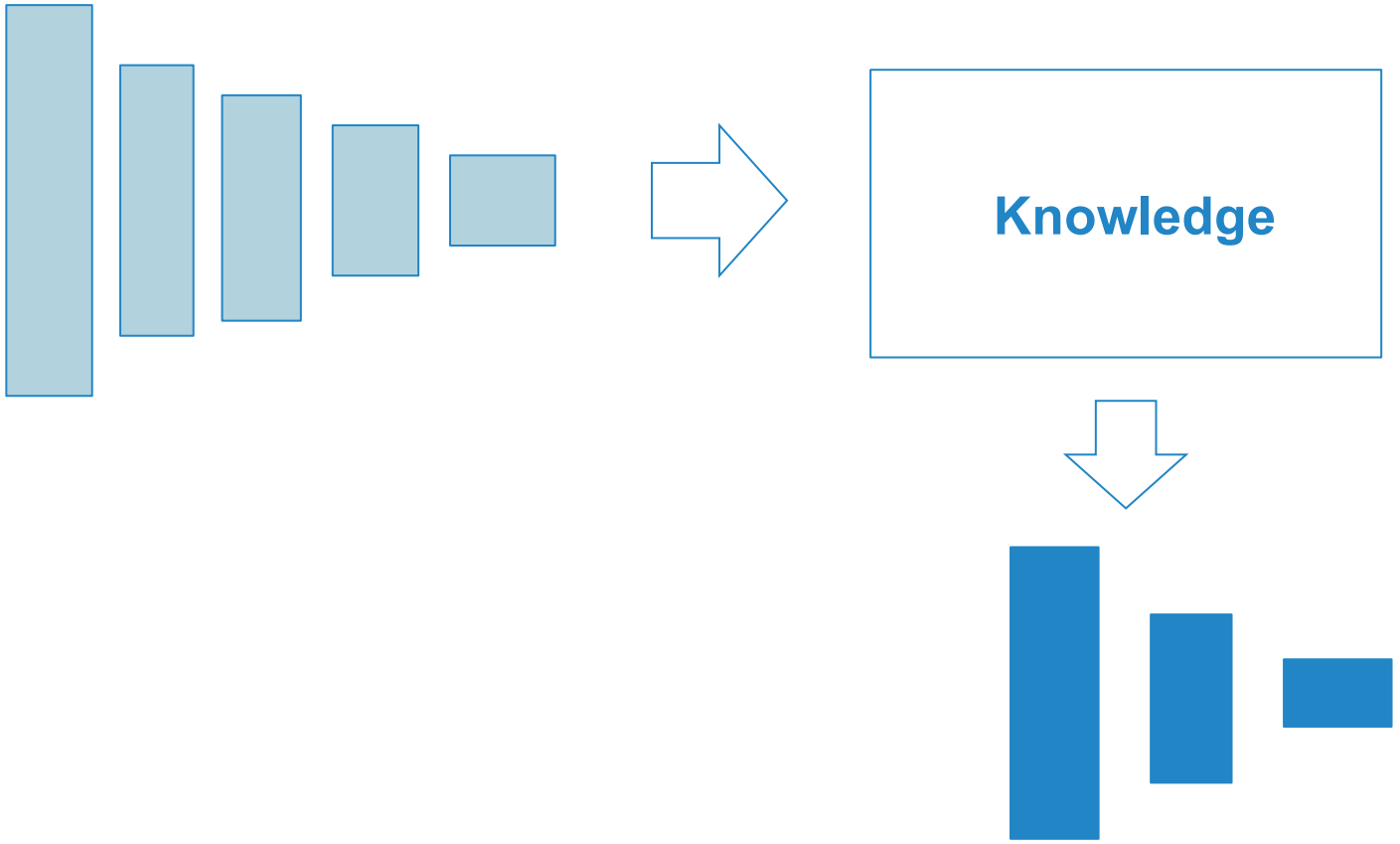
Abstract

- 教師モデルの知識を生徒モデルへ移すKnowledge Distillationを同じアーキテクチャ間で行ったところ，教師モデルの性能を上回る生徒モデルの学習に成功したという研究

- ICML2018採択論文[1]
  - pmlr: http://proceedings.mlr.press/v80/furlanello18a/furlanello18a.pdf
  - arxiv: https://arxiv.org/pdf/1805.04770.pdf

Knowledge Distillation

- Hinton先生らによって提案[2]
- 大規模な既存ネットワークの知識を小さなネットワークに転移させる



Knowledge Distillationのモチベーション

Ridge-i

Large Networks

- 精度が高い
- 表現力が高い
- パラメータ数が多い
- 計算量が大きい

Small Networks

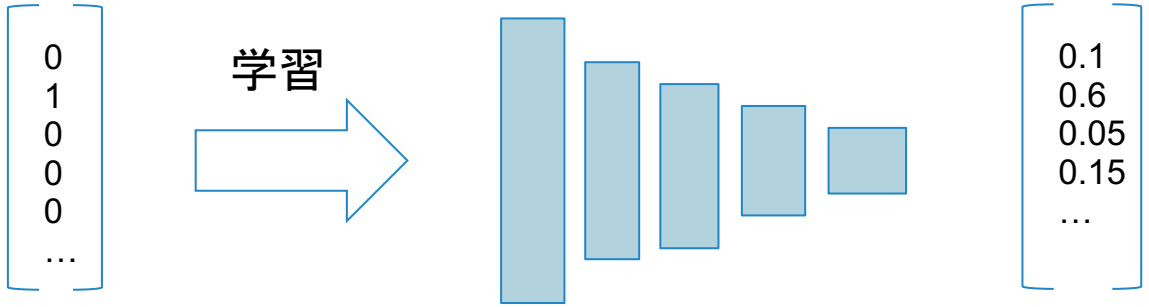
- 精度が低い
- 表現力が低い
- パラメータ数が少ない
- 計算量が小さい

推論時には，モデルの精度だけではなくモデルサイズや処理速度も重要となる
→ 学習時には複雑なモデルで学習し，推論時には軽量なモデルをデプロイしたい

Knowledge Distillationの概要

- NNの学習: パラメータθ，入力xをとる関数$f(x,\theta)$について任意の損失関数を最小化したい
  - $\theta_1^* = argmin_{\theta_1} L\big(y, f(x,\theta_1)\big)$

- 教師モデルの出力分布$f(x,\theta_1^*)$が生徒モデル$f(x,\theta_2^*)$の学習に有用な情報を提供するはず
  - $L(f(x,\theta_1^*), f(x,\theta_2^*))$



ラベル付きデータ　入力画像は猫

教師ネットワーク

- 入力画像は猫
- でもトラにも似てる
- 犬っぽくもある…？

Born-Again Neural Networks (BAN)

- BANでは教師モデルと生徒モデルが同じアーキテクチャを持つケースについて取り組む
  - 同じアーキテクチャ間で知識蒸留を行うとどうなる？

- 逐次的な知識蒸留におけるk番目のモデルの学習

$$\mathrm{L}\left(f\left(x, argmin_{\theta_{k-1}} L\left(f(x, \theta_{k-1})\right)\right), f(x, \theta_k)\right)$$

k-1番目のモデルの出力　　k番目のモデルの出力

- 複数世代の予測の平均化によるアンサンブル学習

$$\hat{f}^{\mathrm{k}}(\mathrm{x}) = \sum_{i=1}^{k} \frac{f(x, \theta_i)}{k}$$

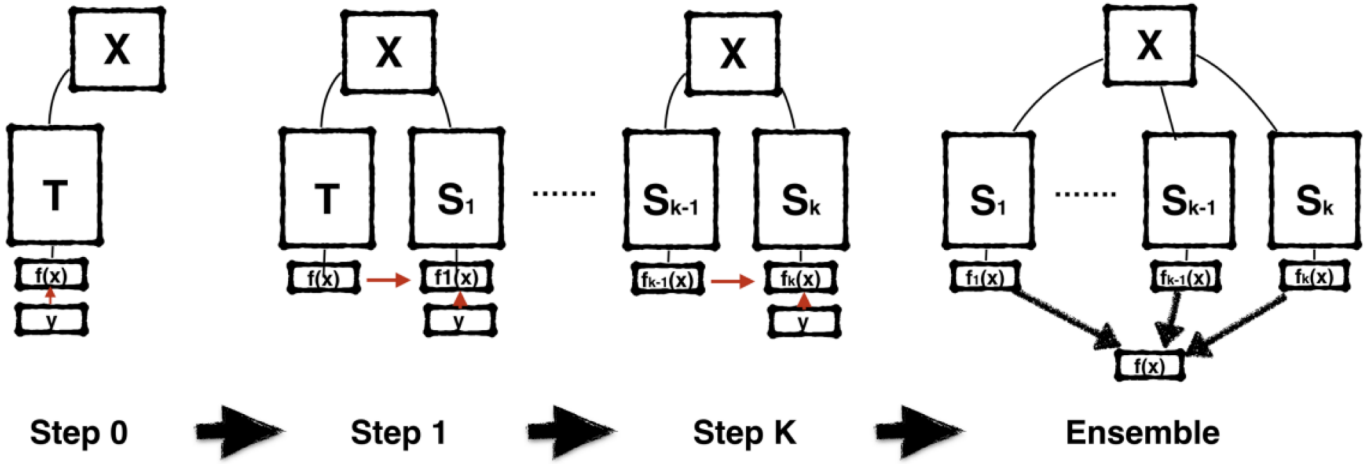
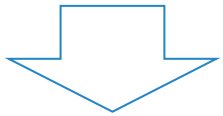


Figure 1. **Graphical representation of the BAN training procedure**: during the first step the teacher model T is trained from the labels Y. Then, at each consecutive step, a new identical model is initialized from a different random seed and trained from the supervision of the earlier generation. At the end of the procedure, additional gains can be achieved with an ensemble of multiple students generations.

Dark Knowledge Under the Light

- 仮説1: Knowledge Distillationはカテゴリ間の類似度を学習している(Hinton et al., 2015)
- 仮説2: 蒸留中と普通の教師付き学習中の正しいクラスに対応する勾配の比較から妥当な説明が得られるのでは



知識蒸留は、正解ラベルの信頼度に対応する重要度重み付けに似ている可能性がある

Ridge-i

- CIFAR-10 Image Classification
  - 全てのケースで精度が向上するわけではないらしい

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

Experimental Results

- CIFAR-100 Image Classification
  - 複数世代のアンサンブルによる精度向上が大きい
  - CIFAR-10では精度が上がらなかった条件でも精度が向上している
    - (タスクの難易度による？)

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**

Conclusion and Discussion

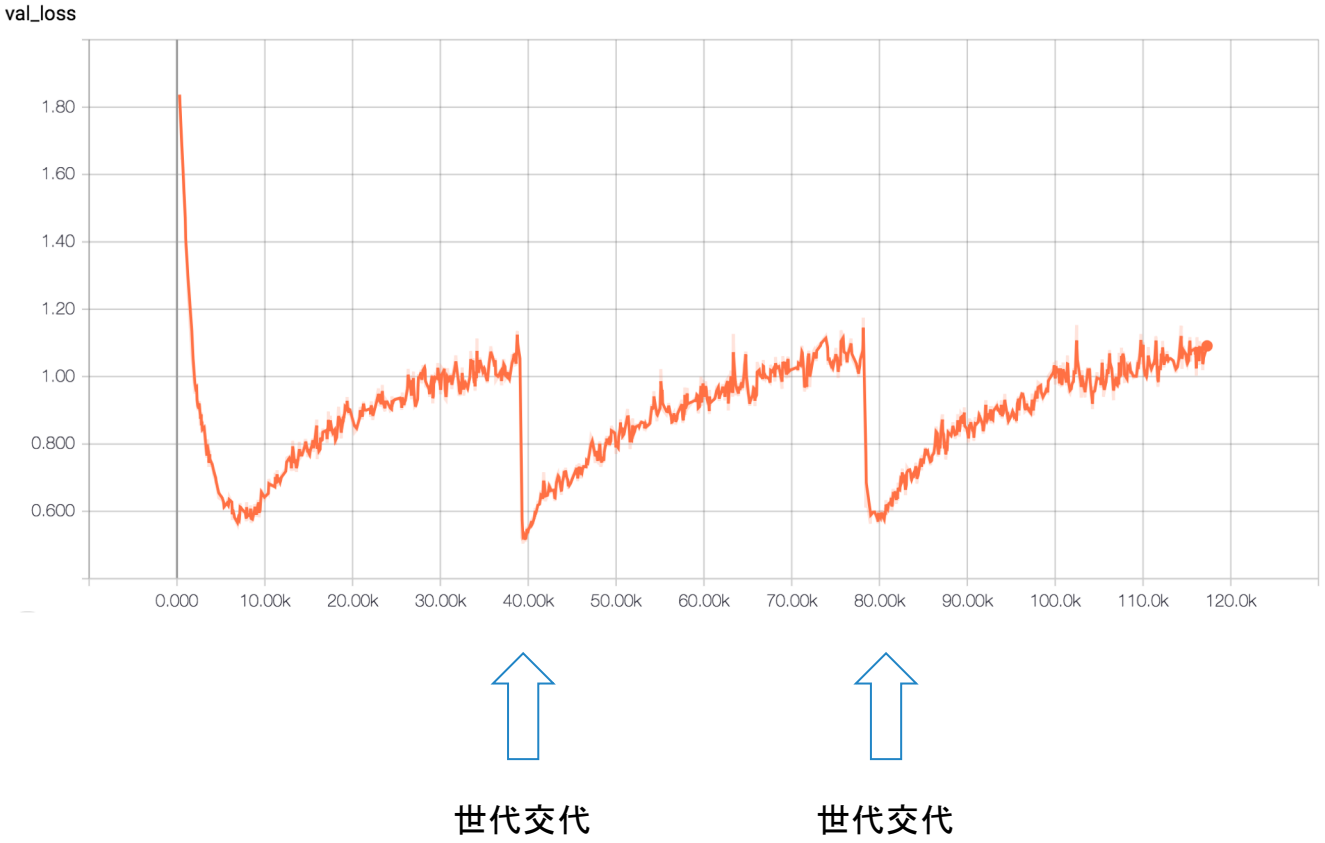
- 同じアーキテクチャ間でのKnowledge Distillationによってモデルの性能向上を達成した
- 蒸留中の勾配と一般的な学習中の勾配の比較から，Knowledge Distillationが学習している知識についての考察が得られた
- 単一の知識蒸留では精度向上が見られないケースがあった一方で，複数世代のアンサンブルをした場合は多くのケースで精度向上を達成できている

おまけ

- 実装してみました
  - Pytorch0.4
  - https://github.com/nocotan/born_again_neuralnet

おまけ(実験)

- データセット: CIFAR-10
  - Data augmentationは無し
- ベースモデル: ResNet-50



Model	Accuracy
第1世代(Baseline)	81.40
第2世代	83.99
第3世代	81.87
Ensemble(1, 2)	84.17
Ensemble(1, 2, 3)	**84.60**

References

1. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L.&Anandkumar, A.. (2018). Born-Again Neural Networks.*Proceedings of the 35th International Conference on Machine Learning, in PMLR*80:1602-1611
2. Hinton, G., Vinyals, O.,&Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.