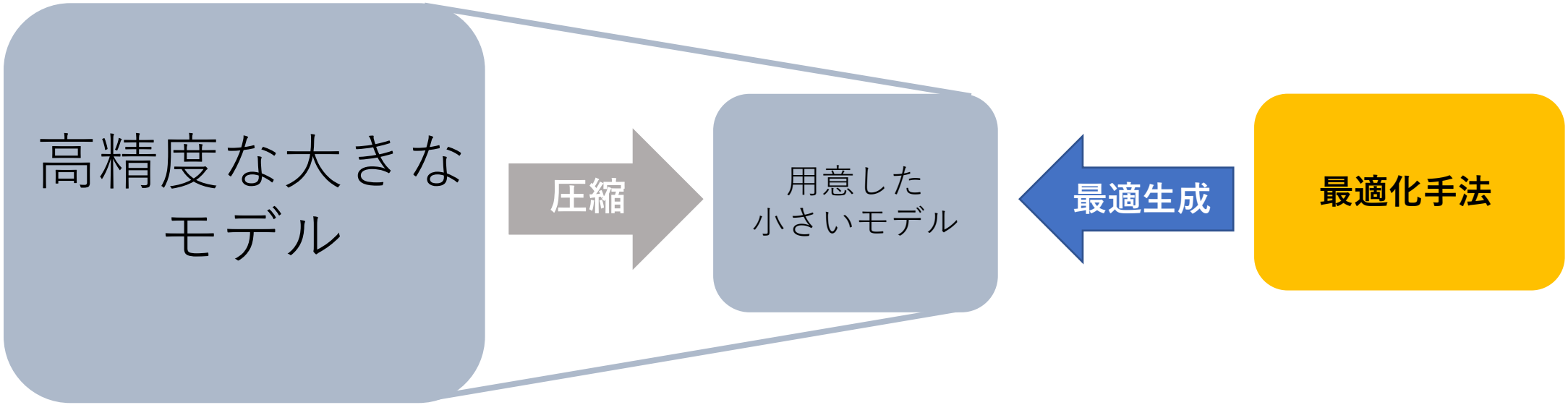
蒸留による
生徒モデルの最適化手法の提案

d-hacks B3 bekku

親：oza

概要

最適化手法を提案し、「深層学習の高精度なモデル」を図のように、圧縮する上で用意する小さなモデルの最適化を実現した。



背景①：高精度なモデルの推論の計算時間

高精度なモデルは、大きく複数のモデルのアンサンブルを用いたりし、

計算量が多くなり、推論の計算時間が長くなることが多い。

様々な機器に実装する場合、推論の計算時間が長過ぎる状態では実用的ではなくなってしまう場合がある。

高速化が求められる

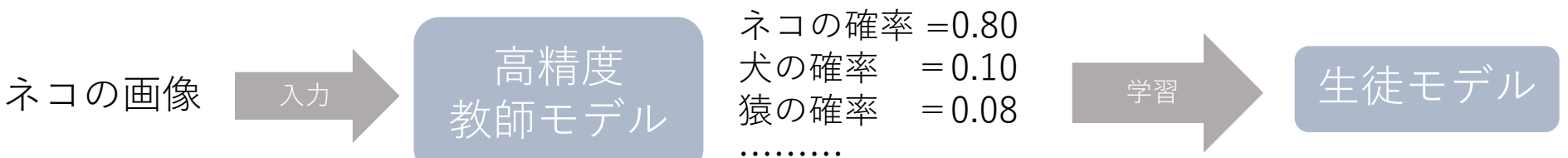
背景②：蒸留による高速化

蒸留とは、モデルを圧縮することで高速化を実現する技術の一つである。
蒸留を用いることで、精度の良いモデルを圧縮させ計算速度を向上させることができる。
生徒モデルの任意性が高く、生徒モデルが**実現できる速度と精度の幅は広い**。

蒸留手法の詳細

教師モデルの誤りも含めた出力と生徒モデルの出力の誤差がなくなるように学習

分類問題の例



同じような結果も出せる
モデルが小さいため、推論が速い

背景③：蒸留の問題

<u>生徒モデルの任意性の高さ故</u>

・モデルが小さいと、高速になるが精度が劣る。

・モデルが大きいと、精度は維持されるが高速化が望めない。

結果

速度・精度の向上の両立は難しいため
生徒モデルの最適がわかりにくい。

関連研究

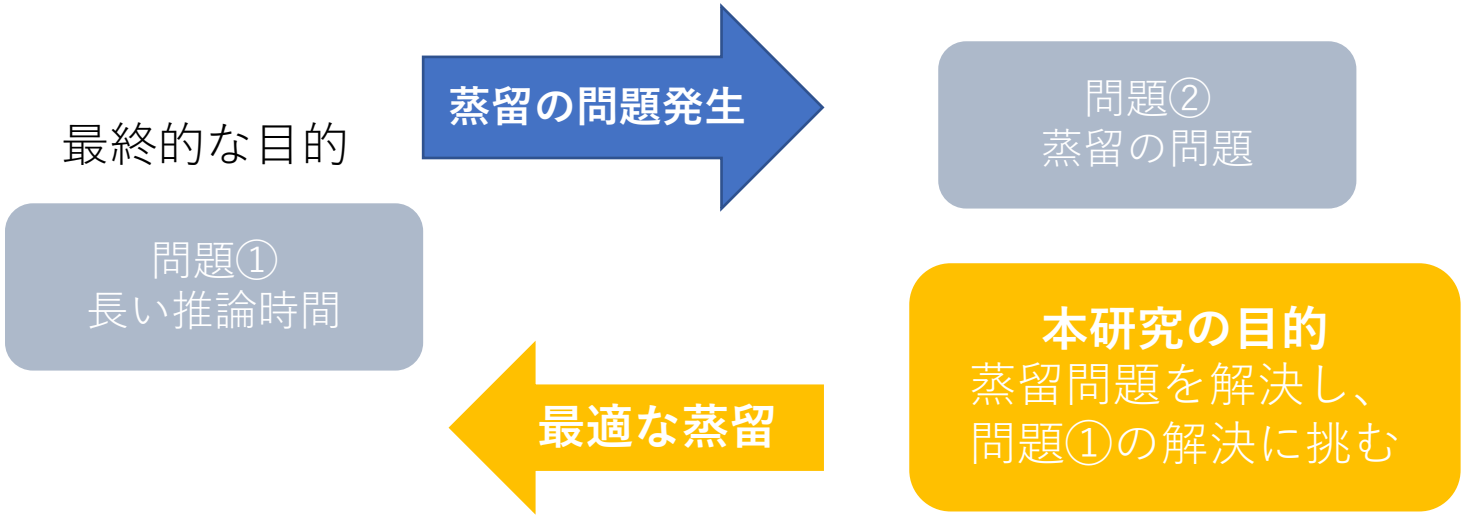
Distilling the Knowledge in a Neural Network
- 温度付きSoftMaxの蒸留学習の提案

関連研究でも、理論的に適した生徒モデルの検討は行っていない。

[1] HINTON, G., VINYALS, O., AND DEAN, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).

本研究の目的

最終的な目的は、蒸留の問題を解決し、高精度な大きなモデルに適用させ高速化を行うことである。



そのため、蒸留の問題の解決を優先し

「任意性の高い生徒モデルを最適化させる」

を本研究の目的とする。

手法の流れ①

推論時間と精度に対して、**評価関数を作り点数をつける。**
その合計値が最も高い生徒モデルを最適な生徒モデルと考える。

生徒モデル①

精度：90点、速度10点
→合計：100点

生徒モデル②

精度：80点、速度70点
→合計：**150点**

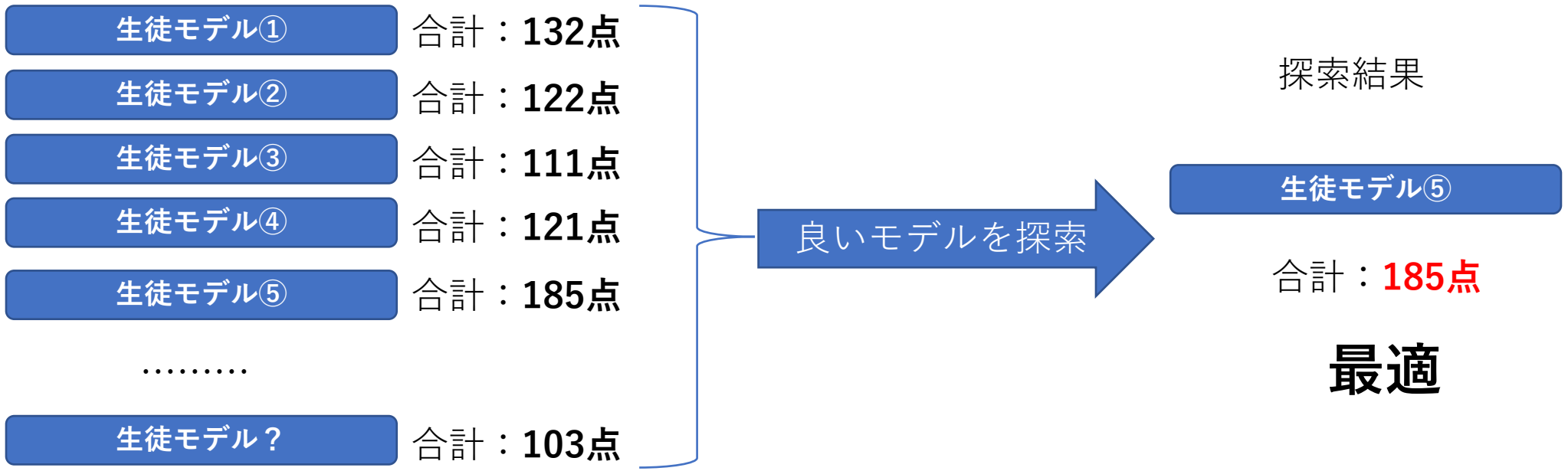
生徒モデル③

精度：50点、速度90点
→合計：140点

生徒モデル②が最適

手法の流れ②

①の評価による点数の合計値が最大となるようなモデルを
組み合わせ最適化問題に適した、**遺伝的アルゴリズム**を用いて探索



評価関数の説明

推論(計算)時間の正規化した評価関数(推論時間が長いほど0に近づき、短いほどr_tに近づく)

$$j_t = (r_t) \times \frac{t_{max} - t}{t_{max} - t_{min}} \ (0 \leq j_t \leq r_t) \ (0 \leq r_t \leq 1)$$

t_{max}: 教師モデルの推論時間、t_{min}： 生徒モデルの推論時間の最小、t $(t_{min} \leq t \leq t_{max})$(範囲外は0or r_t): 生徒モデルの推論時間

精度の正規化した評価関数

$$j_a = (1 - r_t) \times \frac{a - a_{min}}{a_{max} - a_{min}} \ (0 \leq j_a \leq 1 - r_t)$$

a_{max}： 教師モデルの推論精度、a_{min}： 生徒モデルの推論精度の最小、a $(a_{min} \leq a \leq a_{max})$(範囲外は0or $1 - r_t$): 生徒モデルの推論精度
※a_{min}、t_{min}は状況に応じて設定可。

速度と精度の評価関数

$$J(a, t) = j_a + j_t \ (0 \leq J \leq 1)$$

Jが最大の時、最適と定義する。

評価関数の説明

速度と精度の評価関数は、推論時間の最大値r_tを調節することで評価の比重を変更可能

速度を重要視する割合：r_t　　精度を重要視する：$1-r_t$

合計値に対する比重がそれぞれr_t：$1-r_t$となる。

$r_t > 1-r_t$場合、速度を重視した評価を行える。

推論時間や精度に重視するような場合の最適も評価

遺伝的アルゴリズムの説明

NNの例(ノードと層の最適化)

① n世代

[0,1024,2048,128,128]
[0,128,128,128,128]
[0,512,4,2,128]
……
[0,0,0,0,128]

評価値の高いものを選択

② 次世代個体候補

[0,1024,2048,128,128]
[0,128,128,128,128]
[0,1024,2048,128,128]
……
[0,1024,2048,128,128]

交叉
突然変異

③ 次世代作成

[0,1024,2048,128,128]
[0,128,128,128,128]
[0,1024,2048,128,128]
……
[0,1024,2048,1024,128]

上記の①～③の操作を繰り返す。

ランダムに数個取り出し、
その中で**評価値の高いものを次世代の候補とする。**

固体内の遺伝子を**2点間で交叉**する。
一定の確率で、**値にランダム性を持たせる。**

実験の設定や環境

- データセット：Cifar10
- 言語　　：Python
- GPU　　：Google Colab上での「Tesla P100-PCIE-16GB」
- CPU　　：　MacBook Pro (13-inch, 2017, Two Thunderbolt 3 ports)[オプション変更なし] 上のjupyter notebook
- 教師モデル　：VGG16 [2]
- 生徒モデル　：CNN(VGG形式)
- 蒸留手法　　：温度付きSoftMax(温度=10)
- 蒸留epoch数：10
- 評価関数　　：　$t_{min} = 0$、$a_{min} = 0$

遺伝的アルゴリズム

- 世代数　　：20
- 各世代個体数　：30
- 交叉関数　　：2点交差(deap.tools.cxTwoPoint)
- 変異関数　　：　条件の整数で置き換える(deap.tools.mutUniformInt)
- 探索パラメータ：層数、フィルタ数

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. of ICLR, 2015

実験：教師モデル→CNN

VGGを参考とした、生徒モデルの生成方法

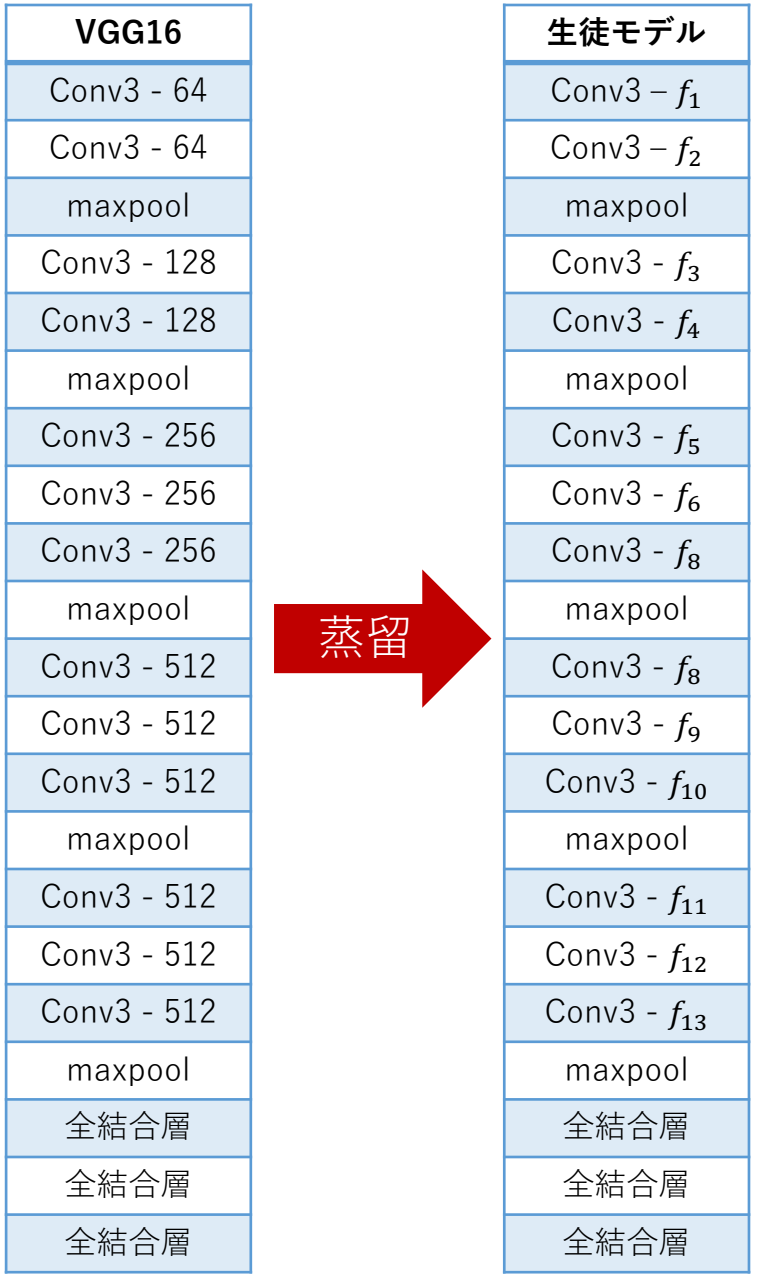
$[f_1, f_2, f_3, , , f_{13}]$を入力として(以下フィルターリスト)、 f_iが各畳み込み層iの**フィルター数**とする。f_iは0、16、32、64、128、256、512のいずれかを取る。また、f_iが0の場合、**層としてみなさない。**

遺伝的アルゴリズムによって、最適モデルの探索

① 手法の有効性の検証 ー 比重１：１$(\boldsymbol{r_t = 0.5})$

② 推論時間と精度の比重を変えた探索の有効性の検証

※右図のConv3 - 64は、畳み込み層のフィルターサイズ3、フィルター数64個を意味する



結果①：手法の有効性の検証($r_t = 0.5$)

	教師モデル	GA探索モデル $r_t = 0.5$で探索	ランダムモデル
GPU上の評価値	×	**0.759**	0.712
精度 (test accuracy)	0.862	**0.820**	0.795
GPU推論時間※(1)	3.498	**2.514**	2.741
CPU推論時間	105.717	**22.219**	35.158

結果①
決め打ちでフィルター数を決めた他のモデルと比べて、GA探索モデルは速度と精度の評価値が優れたものとなった。
教師モデルと比べて、精度は約4%減っているが、GPU上での推論時間は約1.39倍
CPU上での推論時間は約4.75倍となっており高速化も十分行えている。

※(1)この推論時間は、Cifar10のtestデータ10000枚をbatch=128で全て推論する時間。(正確には、時間の値に少し揺らぎが生じるため、20回の平均値)

結果②：推論時間と精度の比重を変えた探索の有効性の検証

		同重視	速度重視	精度重視
	教師モデル	GA探索モデル① $r_t = 0.5$	GA探索モデル② $r_t = 0.75$	GA探索モデル③ $r_t = 0.25$
精度(test accuracy)	0.862	**0.820**	0.774	0.819
GPU推論時間	3.498	2.514	**2.503**	2.581
CPU推論時間	105.717	22.219	**18.165**	33.573

結果②

$r_t = 0.75$

速度に比重を置いたGA探索モデル②は、他のGA探索モデルよりも推論時間において優れている。
教師モデルと比べてGPU上での推論時間は約1.4倍、CPU上での推論時間は約5.82倍となった。

$r_t = 0.25$

精度に比重を置いたGA探索モデル③は、 精度の面において、GA探索モデル①を超えることができなかった。

結果からの考察

- 提案手法により、生徒モデルの適したモデルの探索・生成に成功したと考えられる。

- タスクやデータセットが異なる場合でも、提案手法は適用できるため、手法の汎用性は高く期待できる。

- 結果②より精度に比重を置いたモデルの精度が劣ってしまったのは、遺伝的アルゴリズムは近似的に最適を見つけるアルゴリズムであるため、世代数や世代個体数の設定値が小さかったことが原因だと考えられる。

本研究のまとめ

- 蒸留による、生徒モデルの任意性を問題視
- 速度(推論時間)と精度面の最適化を評価する評価関数の提案
- 最適化における、評価関数と遺伝的アルゴリズムによる探索の有効性の検証

参考文献

[1] HINTON, G., VINYALS, O., AND DEAN, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. of ICLR, 2015