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

Abstract

近年、高精度なモデルを様々な機器への実用化が期待されている。実用化する上で、高精度なモデルの推論の計算時間が長いことが問題となることが多く、高速化が求められている。推論の計算時間を削減する手法の一つとして、蒸留がある。蒸留は高精度な教師モデルの知識を小さい生徒モデルに学習させる技術である。しかし、生徒モデルの任意性は非常に高く速度と精度を両立できるようなモデルを見つけることは困難である。そのため本研究では、速度と精度を評価する評価関数を定義し、遺伝的アルゴリズムを用いて、速度と精度を両立した最適な生徒モデルの探索手法を提案し、最適な生徒モデルの探索手法の検証実験を行う。

1． 背景

ここ数年で深層学習の分野は大きく発展し、様々な領域のデータを扱ったタスクに対して、人間を超える精度を示す高精度な深層学習モデルが出現している。現在では、人間を超えるような精度を示す高精度なモデルが様々な機器への実用化が期待されている。しかし、高精度のようなモデルは深く大きく推論時間が非常に長いことがあり、実際に IoT 等の様々な機器に実装しても、推論の速度が求められるような状況では実用的ではない場合がある。したがって、高精度なモデルの高速化が求められている。

2． 課題

高精度なモデルの高速化手法の一つとして、モデルの圧縮を実現する蒸留という手法がある。蒸留は精度の良い大きな教師モデルの知識を小さな生徒モデルに学習させることで圧縮を実現し、精度を保ちながら推論時の計算速度を向上させる。また、生徒モデルの任意性は高く、生徒モデルが実現できる推論時間と精度の幅は非常に広いと考えられる。したがって、様々な制約をもつ機器に高精度なモデルを実装するための高速化においては、高い柔軟性を持つことから蒸留は非常に優れていると考えられる。しかし、生徒モデルの任意性が高いことから、教師モデルやデータセット、条件によっては最適な生徒モデルは異なる。また、速度と精度はトレースオフのような関係を持ち、速度や精度においての最適な生徒モデルの選択というものは困難な作業となる。

3． 目的

本研究では、課題で述べたような蒸留の生徒モデルの任意性の高さを問題と設定し、任意性の高い生徒モデルを最適化させることとする。

4． 手法

推論時間と精度に対して、それぞれの正規化した評価関数を定義し、その和が最大となる生徒モデルを最適な生徒モデルとする。

$$j_t(t) = \begin{cases} r_t & t < t_{min} \\ \frac{r_t(t_{max}-t+1)}{t_{max}-t_{min}} & t_{min} \leq t \leq t_{max} \\ 0 & t > t_{max} \end{cases} \quad (1)$$

0 から 1 の範囲を取る r_t を最大値とする関数 $j_t(t)$ を推論時間の評価関数と定義する。推論時間の評価関数は、ある条件での生徒モデルの最大の推論時間 t_{max} と、ある条件での生徒モデルの最小の推論時間 t_{min}、評価する生徒モデルの推論時間 t を引数とし正規化を行うことで、生徒モデルの推論時間を 0 から r_t の範囲で評価し評価値を出力する。推論時間の評価値は、評価する生徒モデルの推論時間が設定した最小値に近づくほど r_t に近づき、設定した最大値に近づくほど 0 に近づく。

また、設定した最大の推論時間 t_{max} を超えた場合、評価関数 $j_t(t)$ は 0 を出力し、設定した最小の推論時間 t_{min} を下回った場合は r_t 出力する。

$$j_a(a) = \begin{cases} 0 & a < a_{min} \\ \frac{(1-r_t)(a-a_{min})}{a_{max}-a_{min}} & a_{min} \leq a \leq a_{max} \\ 1-r_t & a > a_{max} \end{cases} \quad (2)$$

0 から 1 の範囲を取る $1-r_t$ を最大値とする関数 $j_a(a)$ を精度の評価関数と定義する。精度の評価関数は、ある条件での生徒モデルの最大の精度 a_{max} と、ある条件での生徒モデルの最小の精度 a_{min}、評価する生徒モデルの精度 a を引数とし正規化を行うことで、生徒モデルの精度を 0 から $1-r_t$ の範囲で評価し評価値を出力する。精度の評価値は、評価する生徒モデルの精度が設定した最小値に近づくほど 0 に近づき、設定した最大値に近づくほど r_t に近づく。

また、設定した最大の精度 a_{max} を超えた場合、評価関数 $j_a(a)$ は r_t を出力し、設定した最小の精度 a_{min} を下回った場合は 0 出力する。

$$J(a,t) = j_t(t) + j_a(a) \quad (3)$$

上記の述べた精度の評価関数と推論時間の評価関数の和を $J(a,t)$ として、速度と精度の評価関数とする。速度と精度の評価関数 $J(a,t)$ が最大の時、最適なモデルとする。この速度と精度の評価関数 $J(a,t)$ は、r_t を調節することで、推論時間の評価関数 $j_t(t)$ の最大値と精度の評価関数 $j_a(a)$ の最大値を変化させられるため、速度と精度の評価関数 $J(a,t)$ へのそれぞれの比重を変更可能であることから、速度や精度に重視するような場合の最適の評価も行えると考えられる。

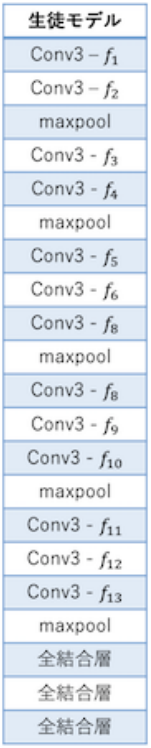

Figure 1: 生徒モデル構成図

生徒モデル生成関数を VGG16 の各畳み込み層のフィルター数を表すリスト ($[f_1, f_2, f_3, , , f_{13}]$) を引数として、Figure 1 のように生徒モデルを生成する。f_i は 0、16、32、64、128、256、512 のいずれかを取り、f_i が 0 の場合、層としてみなさないものとする。

あるフィルター数を表すリストによって生成された生徒モデルを速度と精度の評価関数 $J(a,t)$ によって評価を行い、最大の評価値を出力するフィルター数のリスト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を行う。

5． 実験

まず、評価関数と遺伝的アルゴリズムによる手法の有効性の検証実験を行う。精度の評価関数の最大値である r_t を 0.5 とし、速度と精度の評価関数の比重を 1 対 1 となる生徒モデルを遺伝的アルゴリズムによって探索を行い、理論的な探索や検討を行わず決め打ちで生成した生徒モデルとの評価値を比較する。

次に、速度と精度の比重を変えた探索の有効性についての検証実験を行う。精度の評価関数の最大値である r_t を 0.25、0.5、0.75 として設定した様々な比重を表す評価関数による生徒モデルの探索を遺伝的アルゴリズムによって行い、それぞれの探索モデルの評価値を比較する。

6． 実験結果

6.1 評価関数と遺伝的アルゴリズムによる手法の有効性の検証実験の結果

Table 1: 手法の有効性の検証実験

x	教師モデル	GA 探索	ランダム
評価値	x	**0.759**	0.712
精度	0.862	**0.820**	0.795
GPU 速度	3.498	**2.514**	2.741
CPU 速度	105.717	**22.219**	35.158

GA 探索モデルというのは、今回の提案した遺伝的アルゴリズムを用いて探索したモデルである。GA 探索モデルのフィルター数のリストは、[32, 32, 64, 32, 0, 0, 128, 0, 64, 256, 64, 0, 0] である。ランダムモデルというのは、理論的な探索や検討を行わず、各層のフィルター数を決め打ちで決めたモデルである。ランダムモデルのフィルター数のリストは [64, 32, 256, 16, 64, 16, 64, 512, 64, 32, 64, 256, 64] である。GA 探索モデルは、決め打ちでフィルター数を決めたランダムモデルと比べて、推論時間と精度、速度と精度の全ての評価値が優れたものとなった。教師モデルと比べて、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。

6.2 推論時間と精度の比重を変えた探索の有効性についての検証実験の結果

Table 2: 手法の有効性の検証実験

x	教師	GA ①	GA ②	GA ③
精度	0.862	**0.820**	0.774	0.819
GPU 速度	3.498	2.514	**2.503**	2.581
CPU 速度	105.717	22.219	**18.165**	33.573

GA ①が速度と精度の比重が 1 対 1 で探索を行なったモデルであり、GA ②が速度に比重を置いた探索を行なったモデルであり、GA ③が精度に比重を置いた探索を行なったモデルである。速度に比重を置いたモデルのフィルター数のリストは、[64,16,16,16,128,32,0,0,0,32,64,0,64] である。精度に比重を置いたモデルのフィルター数のリストは、[64, 32, 32, 64, 0, 256, 0, 64, 64, 0, 0,64,512] である。速度に比重を置いた GA ②のモデルは、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。しかし、精度に比重を置いた GA ③のモデルは、精度の面において、GA ①のモデルを超えることはなかった。

7． 考察

手法の有効性の検証から、提案手法によって最適な生徒モデルの探索と生成に成功したと考えられる。また、CNN 以外のモデルに対しても、評価関数と遺伝的アルゴリズムによるパラメータ探索は行える。したがって、タスクやデータセットが異なる場合でも、本研究の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と考えられる。速度と精度の比重を変えた探索の有効性についての検証実験より、精度に比重を置いたモデルの精度が速度と精度の比重を 1 対 1 とした GA 探索モデルに劣った。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いと 1 世代目の初期値に依存してしまい最適に近づけない場合がある。したがって、本研究の実験での世代数が 20、世代個体数が 30 という設定値が小さかったことが原因で最適なモデルを探索できなかったと考える。

8． まとめ

蒸留による、生徒モデルの任意性を問題視し、生徒モデルの速度と精度面の最適を評価する評価関数の提案を行い、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を行なった。検証実験の結果から、手法の有効性は示されたと考えられる。

9． 今後の展望

本研究の検証実験により、設定した世代数や世代個体数の値が最適を見つける上では小さいと考えられるため、実際にこの手法を適用する場合は設定値を十分大きくする必要がある。提案手法は、モデルの探索だけでなく生成も行うため、高精度な大きなモデルを様々な機器に適するモデルに自動で圧縮するシステムとして実用できる可能性がある。本研究の提案手法を用いた上記のシステムによって、様々な状況下の IoT などの小型端末に精度と速度において最適なモデルを気軽に搭載することが期待できる。

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