論文 / 著書情報
Article / Book Information

論題	KL統計量に基づくニューラルネットワークのプルーニング
Title	
著者	町田 兼梧, 井上 中順, 篠田 浩一
Author(s)	Kengo Machida, Nakamasa Inoue, Koichi Shinoda
出典	第22回 画像の認識・理解シンポジウム, , ,
Citation	, , ,
発行日 / Pub. date	2019, 7

KL 統計量に基づくニューラルネットワークのプルーニング

町田 兼梧[1,a)]　井上 中順[1,b)]　篠田 浩一[1,c)]

概要

本研究では，学習済みの多クラス分類器から分類精度向上への貢献が大きいノードを優先的に抽出する KL 平均，KL 分散を提案し，プルーニングに適用する．各提案手法は，ネットワークの各ノードに関する活性度の分布から，より多くのクラスに対して特徴的な反応を示すノードが重要であると仮定して重要度を定義するものである．評価実験では MNIST データセットを用いて従来の手法に基づくプルーニング方法と比較した．提案手法は約 4 割のノードを，精度をほぼ落とすことなく削減できることを示した．また，ノードの削減量が 75%の段階で，提案の指標に基づくプルーニングは従来の相互情報量，KL 最大値に基づくプルーニングを行ったモデルと比較してそれぞれ認識精度が 2.1%，1.8%高くなることを示した．

1. はじめに

近年ディープニューラルネットワーク (DNN) の高い性能が注目され，画像・動画処理や音声認識などに広く応用されている．DNN の構成要素であるノードを複数個取り除いた時に分類精度に与える影響はノードの組合せによって異なり，取り除かれた時に分類器の精度に与える影響が大きいものをより重要度の高いノードと定義する．

学習済みのネットワークのパラメータを剪定する手法 [4] はプルーニングと呼ばれ，重要度の低いノードを取り除くことで効率的なプルーニングが可能である．実際，情報量基準に基づいたノード重要度の推定では，重要度の低い順にノードを切除することで，プルーニング後でも分類精度が高い水準に保たれることが示されている [1].

本研究では Liu ら [1] と同様に，ノードの重要度推定において各ノードの重要度はそのノードの入出力により決定され，他のノードとは独立であると仮定する．Liu ら [1] は，この仮定のもと，各ノードの活性度分布をクラスごとに算出し，分布間の距離に基づいて，特定のクラスに強く反応するノードを重要とみなす指標を提案している．具体的には，クラス c の活性度の分布を p_c，クラスに依存しない活性度の分布 (平均分布) を μ として，μ から p_c へのカルバック・ライブラー距離の最大値 (KL 最大値) をノードの重要度として定義している．この定義は相互情報量に基づいた定義よりも出力層付近でプルーニングに適用した際の効果が高いことが示されている．しかし，KL 最大値は単一クラスに対する活性度分布と平均分布の関係のみを数値化しているため，複数のクラスに共通した特徴に反応するノードの重要度が低くなる．そのため，類似した複数のクラスが存在する場合，ノードがそれらの分類精度向上に貢献していても重要度が低いと推定され，結果としてプルーニング後の分類精度が低下するという問題点がある．

本論文ではこの課題を解決するために，クラスのサブセットに対する活性度分布に基づいた指標を提案する．提案の指標は各サブセットによる活性度の分布と平均分布間の KL 距離平均および分散を重要度として定義するものである．評価実験では，MNIST データセットで学習された中間層が 2 層の全結合ネットワークからプルーニングを行い，従来の指標である相互情報量および KL 最大値との比較を行った．その結果，ノードの抽出量が全体の 4 分の 1 の段階では，中間層 1 層目からの抽出で 2.1%，中間層 2 層目からの抽出で 0.7%分類精度が改善した．

2. 従来手法

2.1 ノード重要度を測る基準値

特定のクラスへの強い反応を示すノード，あるいはその逆であらゆるクラスに反応することにより多くの情報を持つノードは重要であるという仮定から中間層の重要なノードの持つ性質を解明する研究がなされている．

各クラスへのノードの活性度を数値化するものとして神経科学で用いられてきた Selectivity Index(SI) は，データセットのクラス集合を $\mathcal{C}$ として，それぞれのクラスの活性の平均を $m_1, m_2, ..., m_{|\mathcal{C}|}$ とすると

$$\mathrm{SI}_c = \frac{m_c - m_{\mathcal{C}\setminus c}}{m_c + m_{\mathcal{C}\setminus c}} \tag{1}$$

であらわされる．ただし，

$$m_{\mathcal{C}\setminus c} = \frac{1}{|\mathcal{C}| - 1} \sum_{c' \in \mathcal{C}, c' \neq c} m_{c'} \tag{2}$$

[1] 東京工業大学
[a)] machida@ks.c.titech.ac.jp
[b)] inoue@c.titech.ac.jp
[c)] shinoda@c.titech.ac.jp

とした．$-1 \leq \mathrm{SI}_c \leq 1$ であり，$\mathrm{SI}_c = -1$ はノードがクラス c に全く反応しない，$\mathrm{SI}_c = 1$ はノードがクラス c にのみ反応することを示す．Ari ら [6] は，クラス選択性 (class selectivity)S を

$$\mathrm{S} = \max_c \mathrm{SI}_c \tag{3}$$

と定義し，ノード重要度の指標として利用した．Bolei ら [7] はこの指標を用いて逐次的にノードを切除する手法により重要なノードを抽出している．

また，Liu らは情報理論を用いてノード重要度を数値化をしている [1]．ノードの出力を閾値に従って量子化することにより生成された確率変数を V，クラスの集合 $\mathcal{C}$ として各クラスを $c \in \mathcal{C}$ とすると，相互情報量 $I(V;\mathcal{C})$,KL 情報量 $D_{\mathrm{KL}}(p\|p')$ はそれぞれ

$$I(V;\mathcal{C}) = \sum_{v\in V}\sum_{c\in\mathcal{C}} p(v,c)\log\frac{p(v,c)}{p(v)p(c)} \tag{4}$$

$$D_{\mathrm{KL}}(p\|p') = \sum_v p(v)\log\frac{p(v)}{p'(v)} \tag{5}$$

で定義される．また，$p_c(v) = p(v|c)$ をクラス c における活性度分布，$p(v|c)$ の全クラスに渡る平均 $\mu(v)$ を平均分布と定義する．これらを用いて，従来のノード重要度を測る指標は相互情報量 (式 (4)) の他に KL 情報量 (式 (5)) の全クラスに渡る最大値である KL 最大値が用いられた．KL 最大値 $\mathrm{Max}_{\mathrm{KL}}$ は

$$\mathrm{Max}_{\mathrm{KL}} = \max_c \ D_{\mathrm{KL}}(p_c(v)\|\mu(v)) \tag{6}$$

で表される．以降 KL 情報量を KL 距離と呼ぶ．また，ノード重要度はそのノードの出力のみで決まり，他のノードに左右されないことに注意されたい．

2.2 課題

従来の手法ではクラスの情報を単一クラスのみで扱っていたため複数のクラスにより生成される分布と平均分布との関係性が考慮されていなかった．また，入力層に近い中間層では相互情報量，出力層に近い中間層では KL 最大値が重要なノードを抽出できるという結果であったため，すべての中間層において重要なノードを抽出できる統一的な指標が望ましい．

3. 提案手法

2.2 節での課題を考慮して，複数クラスのサブセットにより生成される分布も考慮に入れた指標を提案する．あるクラスにラベル付けされたデータを入力したときの出力が偏っているほど特徴的な反応であり，重要であると考えられる．これがすべてのクラスにおいて成り立つノードは重要であると考えられ，出力の偏ったノードは各クラスにおける分布と平均分布の KL 距離は大きくなることから，2 つの指標，KL 平均と KL 分散を提案する．

3.1 ノード活性度の分布推定

まず，サンプル x_i とそのクラスラベル $y_i \in \mathcal{C}$ から成るデータセット $\mathcal{D} = \{(x_1,y_1),\cdots,(x_N,y_N)\}$ で学習されたニューラルネットワーク考え，x_i を入力した際の l 層 k 番目のノードの出力を $t_k^{(l)}(x_i)$ とおく．一般に，この出力は活性化関数を適用した直後の値であり，全結合ネットワークを例にとると，

$$t_k^{(l)}(x_i) = \sigma\left(\sum_{k'} w_{k',k}^{(l-1)} t_{k'}^{(l-1)}(x_i) + b_k^{(l)}\right) \tag{7}$$

で表される．ここで，$w_{k',k}^{(l-1)}$ は重み係数，$b_k^{(l)}$ はバイアス係数，$t_{k'}^{(l-1)}(x_i)$ は $l-1$ 層からの入力，σ は活性化関数である．

次に，Liu らの先行研究 [1] と同様に，ノードの出力 $t_k^{(l)}$ を量子化する量子化器 $q : \mathbb{R} \to \mathcal{T}$ を導入すると，l 層 k 番目のクラス c に関するノード活性度の分布 $p_c^{(l,k)}$ は

$$p_c^{(l,k)}(Q_{l,k} = v) = \frac{1}{N_c}\sum_{i=1}^{N} \mathcal{X}[y_i = c, q(t_k^{(l)}(x_i)) = v] \tag{8}$$

と表される．ここで，$Q_{l,k}$ は $t_k^{(l)}$ の量子化結果を表す $\mathcal{T}$ 上の確率変数，N_c はクラスラベルが c であるサンプルの総数，$\mathcal{X}[A]$ は条件 A を満たす時に 1，そうで無い時に 0 をとる指示関数である．また，同様にクラスに依存しないノード活性度の分布を表す平均分布は

$$\mu^{(l,k)}(Q_{l,k} = v) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{X}[q(t_k^{(l)}(x_i)) = v] \tag{9}$$

と表される．以下では Liu らの研究 [1] と条件を揃えるため，$\mathcal{T} = \{0,1\}$ とし，量子化器 Q のしきい値は活性化関数が Sigmoid の時に 0.5，ReLU の時に 0 とする．

本研究では，クラスのサブセットに対する活性度の分布を導入するため，$C_S \subset \mathcal{C}$ に対しても，以下のように分布 p_{C_S} を求める．

$$p_{C_S}^{(l,k)}(Q_{l,k} = v) = \frac{1}{\sum_{c\in C_S} N_c}\sum_{i=1}^{N} \mathcal{X}[y_i \in C_S, q(t_k^{(l)}(x_i)) = v] \tag{10}$$

3.2 KL 平均 (KL mean)

KL 平均はクラスに依存しない平均分布 μ とクラスまたはクラスのサブセットが与えられた際に定まる活性度分布 p_{C_S} の KL 情報量 (KL 距離) を距離とみなして，これらすべてを平均する指標である．クラスにより定まる分布と平均分布との距離が大きい場合，そのクラスの活性は他のクラスによる活性との差が大きく重要な活性であると考えられる．これらをすべてのクラスについて平均することで多くのクラスで差別化された活性を示すノードを抽出する指標であり，l 層 k 番目のノードの KL 平均 $\mathrm{Mean}_{\mathrm{KL}}^{(l,k)}$ を以下で定義する．

$$\mathrm{Mean}_{\mathrm{KL}}^{(l,k)} = \frac{1}{2^{|\mathcal{C}|}} \sum_{C_S \subseteq \mathcal{C}} D_{\mathrm{KL}}(p_{C_S}^{(l,k)} \| \mu^{(l,k)}) \qquad (11)$$

ここで，$D_{\mathrm{KL}}(\cdot\|\cdot)$ はカルバック・ライブラー情報量である．

3.3 KL 分散 (KL variance)

従来の指標である KL 最大値は，KL 距離の最大値を与えるクラスと平均分布との KL 距離で与えられていたため，最大距離を与えるクラス以外の KL 距離も大きくなる場合がある．この場合は他のクラスとの差別化ができない．KL 分散はこの問題を解決し，クラスのサブセット間で活性度分布がどの程度異なるのかを数値化する指標であり，l 層 k 番目のノードの KL 分散 $\mathrm{Var}_{\mathrm{KL}}^{(l,k)}$ は以下のように定義される．

$$\mathrm{Var}_{\mathrm{KL}}^{(l,k)} = \frac{1}{2^{|\mathcal{C}|}} \sum_{C_S \subseteq \mathcal{C}} (D_{\mathrm{KL}}(p_{C_S}^{(l,k)} \| \mu^{(l,k)}) - \mathrm{Mean}_{\mathrm{KL}}^{(l,k)})^2 \qquad (12)$$

4. 実験

4.1 ネットワークの構造と学習時の設定

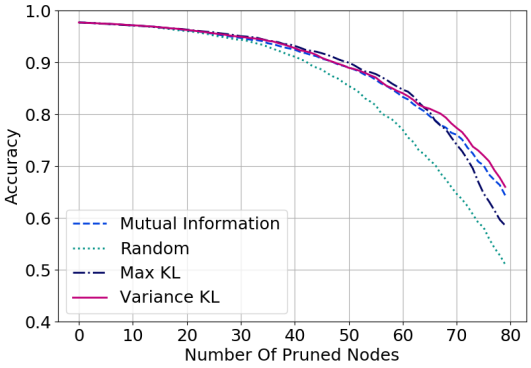
評価実験は，MNIST データセット [5] で，全結合層 2 層のネットワークを用いて実施した．各層それぞれ 100 ノードであり，活性化関数は sigmoid を用いた．最適化関数はクロスエントロピーで，L2 正則化を行っている．最適化手法は Adam[3] を用いた．学習率は 0.001，バッチサイズは 32 として，80 エポックでトレーニングを行っている．

4.2 層ごとの各基準値に従った逐次的なプルーニング

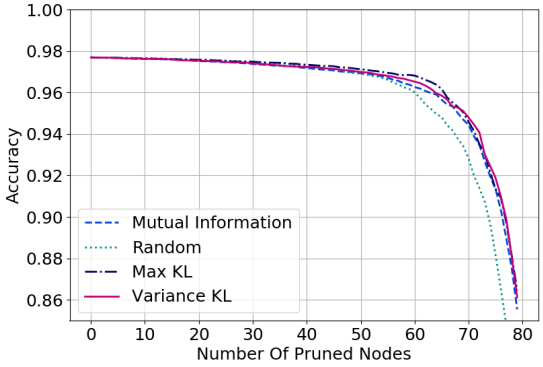
層ごとにノードを逐次的に切除した際の分類精度の変化を図 1，図 2 に示す．KL 平均と KL 分散を，相互情報量，KL 最大値のほか，ランダムに切除した場合と比較した．より正確なデータを得るため別々に訓練を行った 20 個のモデルを用意し，それぞれのモデルに対してノードを逐次的にプルーニングした際の平均精度を報告している．

図 1 では，新たに導入した KL 分散は実線でプロットされている．図 1(a) が中間層 1 層目，図 1(b) が中間層 2 層目であり，各基準値が小さいノードから取り除いた際のテストセットにおける分類精度である．図 1 より，提案された KL 分散の小さい値を示すノードの逐次的な切除は，他の基準値と比較して中間層 1 層目，2 層目どちらに対しても精度が低下しないことが分かる．特に 2 層目からのノードの切除では精度をほぼ落とすことなく 4 割のノードを削減することが可能である．各基準値に従ってノードの切除する数を 60，65，70，75 とした際の精度を表 1 に示す．表 1(a) は図 1(a)，表 1(b) は図 1(b) にそれぞれ対応している．表 1 より，提案された KL 分散は，特に切除するノードの数が 70 を超えてから 1 層目，2 層目ともに他の基準値に基き切除を行ったモデルと比較して精度が高い．

さらに，KL 平均と相互情報量，KL 最大値の比較の図



(a) 1 層目:基準値の小さい順に切除



(b) 2 層目:基準値の小さい順に切除

図 1: 層ごとの逐次的なノード切除．KL 分散と他の基準値の比較．横軸は切除したノードの数，縦軸はテストセットによる精度．

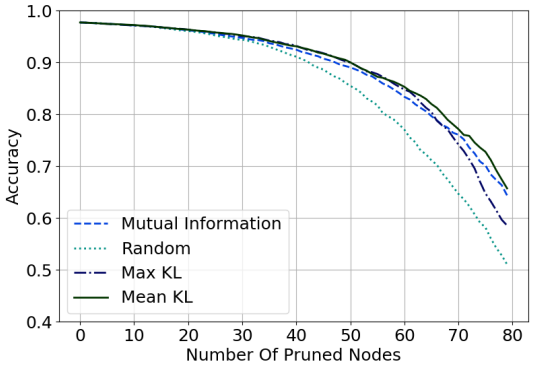
切除数	相互情報量による切除	KL 最大値による切除	KL 分散による切除
60	0.833	**0.847**	0.840
65	0.796	0.804	**0.810**
70	0.760	0.741	**0.773**
75	0.701	0.647	**0.719**

(a) 1 層目:基準値の小さい順に切除

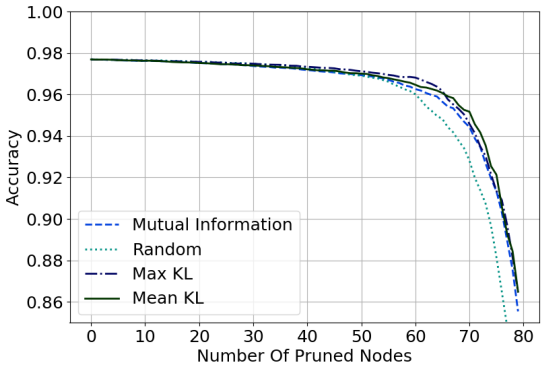
切除数	相互情報量による切除	KL 最大値による切除	KL 分散による切除
60	0.962	**0.968**	0.965
65	0.956	**0.961**	0.958
70	0.944	0.945	**0.947**
75	0.913	0.913	**0.918**

(b) 2 層目:基準値の小さい順に切除

表 1: KL 分散と相互情報量，KL 最大値の比較．1 層目，2 層目から各基準値の小さい順に逐次的なプルーニングを行った際のテストセットにおける精度．一番精度の良いものを太字で示してある．



(a) 1 層目:基準値の小さい順に切除



(b) 2 層目:基準値の小さい順に切除

図 2: 層ごとの逐次的なノード切除．KL 平均と他の基準値の比較．横軸は切除したノードの数，縦軸はテストセットによる精度.

を図 2 に示す．新たに導入した KL 平均は実線でプロットされている．KL 分散の実験と同様に，各基準値に従ってノードの切除する数を 60，65，70，75 とした際の精度を表 2 に示した．表 2(a) は図 2(a)，表 2(b) は図 2(b) にそれぞれ対応している．表 2 より，KL 分散と同様に切除するノードの数が 70 を超えてからは 1 層目，2 層目ともに他の基準値に従ってプルーニングを行ったモデルと比較して精度が高い結果となった.

以上より，特に切除するノードが 7 割を超えてからは，中間層が 2 層のモデルに対して，1 層目，2 層目で重要なノードを抽出する相互情報量，KL 最大値に従ってプルーニングを行ったモデルと比較してそれぞれ 2%，0.5%程度精度が上昇し，全ての層でニューロンの重要度を測る指標として従来のものよりも優れていると結論づけた.

5. まとめと今後の課題

本論文では，分類タスクにおいて重要なノードの特徴を解明し，重要となるノードを順序付ける基準値である KL 平均と KL 分散を提案した．提案された 2 つの指標はともに従来の指標である相互情報量，KL 最大値と比較して，

切除数	相互情報量による切除	KL 最大値による切除	KL 平均による切除
60	0.833	0.847	**0.849**
65	0.796	0.804	**0.818**
70	0.760	0.741	**0.769**
75	0.701	0.647	**0.722**

(a) 1 層目:基準値の小さい順に切除

切除数	相互情報量による切除	KL 最大値による切除	KL 平均による切除
60	0.962	**0.968**	0.965
65	0.956	**0.961**	0.960
70	0.944	0.945	**0.949**
75	0.913	0.913	**0.920**

(b) 2 層目:基準値の小さい順に切除

表 2: KL 平均と相互情報量，KL 最大値の比較．1 層目，2 層目から各基準値の小さい順に逐次的なプルーニングを行った際のテストセットにおける精度．一番精度の良いものを太字で示してある.

いかなる層においても重要度の高いノードの抽出が可能である点で優れている．一方で，切除するノードの数によっては既存の基準値がより適している場合があるため，切除するノードの数とその際に重要になる指標の関係性の解明が今後の課題である.

参考文献

[1] K. Liu, R. A. Amjad, and B. C. Geiger, “Understanding Individual Neuron Importance Using Information Theory,” arXiv preprint arXiv:1804.06679, 2018.

[2] B. Zhou, Y. Sun, D. Bau, and A. Torralba, “Revisiting the Importance of Individual Units in CNNs via Ablation,” arXiv preprint arXiv:1806.02891, 2018.

[3] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” International Conference on Learning Representations (ICLR), 2015.

[4] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” International Conference on Learning Representations (ICLR), 2016.

[5] Y. LeCun, “The MNIST database of handwritten digits,” com/exdb/mnist/, 1998. 30 http://yann. lecun.

[6] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, “On the importance of single directions for generalization,” International Conference on Learning Representations (ICLR), 2018.

[7] B. Zhou, Y. Sun, D. Bau, and A. Torralba, “ Revisiting the Importance of Individual Units in CNNs via Ablation, ” arXiv preprint arXiv:1806.02891, 2018.