Vol.2023-IOT-60 No.21
2023/3/16

DDoS検出ための畳み込みニューラルネットワークの軽量化手法について

趙　彦博[1,a)]　小林　孝史[2,b)]

概要：ネットワークの発展やアクセスするデバイスの急増に伴い，DDoS 攻撃の強度や頻度が高まっており，DDoS 攻撃の緩和や検知が重要となっている．CNN に代表される機械学習アルゴリズムは DDoS 攻撃検知の分野でも応用され，良好な成果を得ている．しかし，CNN はもともと画像領域で使われるニューラルネットワークであり，DDoS 検知に適用する場合，パラメータ数や実行速度が問題となる．本研究では，知識蒸留により CNN モデルを圧縮することで，軽量な DDoS 攻撃検知を実現する．本研究はまず，CNN を用いてネットワークトラフィックの二値分類を行い，DDoS 攻撃の有無を検出する検出率 99.8%を達成し，DDoS 攻撃検出分野における CNN の有効性を検証した．その後，6 つの CNN モデルを選定し，CICDDoS2019 データセットで多目的分類を行い，最適なモデルを教師モデルとして，LSTM を生徒モデルに選定し，知識蒸留を行う．最終的に，LSTM モデル単独での学習より 1.6%高い，86.9%の正解率を達成し，パラメータ数と学習時間を大幅に削減することができた．

キーワード：DDoS 攻撃，攻撃検出，畳み込みニューラルネットワーク，知識蒸留

A Lightweight Method for Convolutional Neural Networks for DDoS Detection

Abstract: With the development of networks and the rapid increase in the number of devices accessing them, the intensity and frequency of DDoS attacks have increased, making DDoS attack mitigation and detection important. Machine learning algorithms such as CNNs have been applied to DDoS attack detection with good results. However, CNNs are neural networks originally used in the image domain, and when applied to DDoS detection, the number of parameters and execution speed are problematic. In this study, we use knowledge distillation to compress CNN models to achieve lightweight DDoS attack detection. First, this research verified the effectiveness of CNNs in the field of DDoS attack detection by achieving a detection rate of 99.8 percent by using CNNs to bi-classify network traffic and detecting the presence of DDoS attacks. We then selected six CNN models and performed multi-objective classification on the CICDDoS2019 dataset, selecting the best model as the teacher model and LSTM as the student model for knowledge distillation. Finally, we achieved an 86.9% correct response rate, 1.6% higher than the LSTM model, and the number of parameters and training time were significantly reduced.

Keywords: DDoS attack, attack detection, convolutional neural networks, knowledge distillation

1. はじめに

ネットワークを攻撃する方法の 1 つに，DDoS 攻撃が存在する．DDoS 攻撃は，分散的にターゲットのネットワークサービスの処理能力を超える大量のデータパケットをターゲットに送信する攻撃であり，分散型サービス妨害攻撃とも呼ばれている．DDoS 攻撃によってネットワークサービスを継続できないといった影響がある．

ネットワークの発展とネットワークにアクセスするデバイスの急増に伴い，攻撃の強度と頻度も増加しており，ネットワークシステムが深刻な危険にさらされている．ネットワークは今日の社会になくてはならないものであり，電力，

[1] 関西大学　総合情報学研究科
Graduate School of Informatics, Kansai University
[2] 関西大学　総合情報学部
Faculty of Informatics, Kansai Uniersity
[a)] k326527@kansai-u.ac.jp
[b)] taka-k@kansai-u.ac.jp

政府，金融などの重要な分野でのネットワークサービスの障害は大きな社会問題を引き起こすことにもなるため，DDoS 攻撃からの保護は非常に重要である．また，攻撃が発生したときに攻撃を緩和することに加えて，DDoS 攻撃を検出することも重要である．

DDoS 攻撃の研究には，攻撃の検出や緩和，攻撃手法の特定など，様々な研究テーマがある．緩和策としては，攻撃対象のインフラの規模を拡大することによって負荷を分散する，流量の制御，攻撃元のブロックなどが考えられる．また，攻撃手法に関しては Botnet，Amplification Attack などがあり，それぞれに対して対策をすることが重要である．

検出には，流量分析，パケット解析，機械学習などの手法が使用されている．トラフィックデータ分析によって，攻撃によって生じる流量の変化を検知し，攻撃を検出する．パケット解析を使用する方法では，送信元 IP アドレスや，送信元のポート番号などを分析し，攻撃を検出することができる．機械学習によっては，学習済みのモデルを用いて，流量データやパケットデータから攻撃を検出することができる．

検出に関しては検出精度を重視する手法と，検出速度を重視する手法がある．検出精度を重視する手法では，攻撃を正確に検出するために，計算資源を多く消費し，また，検出速度を重視する手法では，リアルタイムに攻撃を検出するために，精度は劣るが，低い計算資源で検出できるといった特徴がある．

既存の検出方法の中で機械学習は主要な方法の一つである．現在，マシンラーニングとディープラーニングは DDoS 攻撃検出の分野でうまく適応され，良好な結果を達成しており，特にディープラーニングの CNN は最も高い検出率を示している．しかし，DDoS 検知に使う場合はパラメータ数と実行速度に関して課題がある．

本研究では，CNN を使用して，CICDDoS2019 データセットのトラフィックデータを分類する．また，モデルの圧縮を行うことを目的としている．

2. 関連研究

Aanshi Bhardwaj ら [1] は DNN とオートエンコーダーを融合し，DDoS 検出のために DNN アーキテクチャを適応させ，最終的に NSL-KDD と CICIDS2017 でテストした． 彼らはまずデータセットをオートエンコーダで符号化し，それを DNN に渡した．オートエンコーダーはグリッドサーチとスパース性を利用して，チューニングした．DNN は最新のハイパーパラメータを使用する．最終的に NSL-KDD で 98.43%の精度を達成し，CICIDS2017 データセットで 98.92%の検出率を実現した．

Priyadarshini ら [2] はクラウドコンピューティングとフォグコンピューティング環境のネットワークセキュリティのために, SDN 制御層で DDoS 攻撃を検出できる Long Short Term Memory (LSTM) 深層学習モデルを提案した．時間ベースの逐次データでうまく機能する LSTM モデルは，前のパケットが現在のパケットに与える影響に関する知識を保持するため，ある時間間隔でキャプチャしたネットワークパケットを用いた学習に適している．実験の結果，隠れ層 3 層，ユニット数 128 からなる LSTM 深層学習モデルが適切であることがわかる．SCX 2012 と IDS CTU-13 ボットネットのデータセットで 98.8%の検出率を達成し，hping3 で DDoS 攻撃をシミュレーションして，テストを行った．

Abdullah EmirCil ら [3] は CICDDoS2019 データセットをチューニングし，フィードフォワードニューラルネットワーク（FFN）を運用して DDoS 攻撃検出を行った．彼らは，8 つの静的な特徴（Flow ID，SourceIP，SourcePort，DestinationIP，DestinationPort，Protocol，Timestamp，SimillarHTTP）を消去する同時に，また 9 つのデータはほぼ 0 の特徴（Fwd PSH Flags, Fwd URG Flags, Bwd URG Flags, Fwd Bytes/Bulk Avg, Fwd Packet/Bulk Avg, Fwd Bulk Rate Avg, Bwd Bytes/Bulk Avg, Bwd Packet/Bulk Avg，Bwd Bulk Rate Avg）であったので，これらもデータセットから省いている．二値分類で 99.99%，多目的分類で 94.5%の検出率を達成した．

Faisal Hussain と Syed Ghazanfar Abbas ら [4] は ResNet18 を利用して，DDoS の攻撃検出を実現した．彼らは，CICDDoS2019 のデータセットから 60 個の特徴のみを残して実験を行った．そして，データの極大値を削除し，150 個のデータを 1 セットとして，3 セットを入力画像の 3 チャンネルとして選択し，2500 枚の 150*150*3 の RGB 画像に変換する．実験用 ResNet のハイパーパラメータは，LEARNING RATE は 0.0001，momentum は 0.9，optimizer は SGD，epochs は二値分類 10 回，多目的分類 50 回で実験を行う．二値分類で 99.9%，多目的分類で 87%の検出率を達成した．

Yuelei Xiao ら [5] は, ResNet を DDoS のような簡単な分類に直接適用すると，モデルが複雑になりすぎてオーバーフィットしやすくなることを発見し，ResNet18 ではモデル刈り込みを適用してモデル圧縮を行っている．ResNet18 はもともと 10 層の畳み込み層と 8 層のプーリング層を持っていたが，各残差ブロックから 2 層の畳み込み層のうちの 1 つを取り除き，活性化関数を ReLu から PReLu に変更した．DDoS 検知のためのモデル刈り込みを実施し，NSL-KDD データセットでオーバーサンプリング後のバイナリ検知率 99.5%を達成した．

3. 知識蒸留

蒸留とは，汚れた水を蒸発させ，凝縮冷却してより純粋な水を生成することである．2015 年に Hinton[6] が提案し

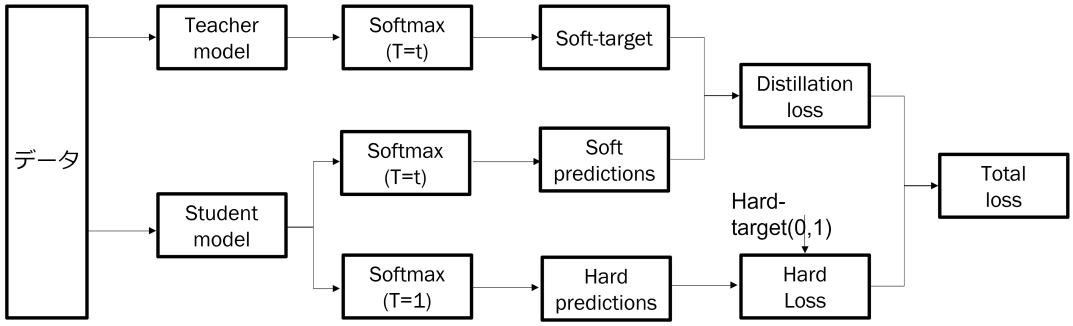

図 1 知識蒸留の流れ
Fig. 1 Flow of knowledge distillation

た知識蒸留（knowledge distillation，KD）は，複雑な知識からより有用な知識を蒸留することを目指している．また複雑な構造と優れた性能を持つ教師モデルで soft-target を生成し，簡単な構造で展開に適した生徒モデルの学習を支援する．Hinton の論文の厳密な数学的導出は複雑なものではないが，主に蒸留ニューラルネットワークの実現可能性を検証している．

知識蒸留の本質は Student Network（生徒モデル，SN）が Teacher Network（教師モデル，TN）の汎化能力を学習することであり，汎化能力の高い TN が既にあるので，TN を使って SN を蒸留訓練する際に，TN の汎化能力を直接 SN に学習させることができる．汎化能力を教える非常に簡単で効率的な方法は，softmax 層が出力するカテゴリの確率を「soft-target」として利用することである．従来の学習プロセス（hard-target）は，入出力関係の最適化である．KD 学習プロセス（soft-target）は，大規模モデルの確率分布である．

活性化関数 softmax は複数の出力値の合計が 1.0 に変換できる関数である．Hinton は softmax に温度関数 T を取り込んで計算する．式（1）は，softmax 関数に温度関数 T を追加したものである．本来の softmax 関数は，$T = 1$ の特殊なケースである．T が大きいほど，softmax の出力確率分布が滑らかになり，その分布のエントロピーが大きくなるため，ネガティブラベルが持つ情報が相対的に増幅され，モデルの学習がネガティブラベルに集中することになる．

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{1}$$

知識蒸留の流れは二段階に分かれている．第 1 段階は TN の学習を行う．第 2 段階は高い T で TN から SN への知識蒸留を行う．

第 2 段階の高温蒸留について説明する．高温蒸留プロセスの目的関数は，distill loss（soft-target に相当）と student loss（hard-target に相当）を重み付けして求めている．

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

式（2）の Loss 関数の第 1 項は式 (3) のようになる：

$$L_{soft} = -\sum_j^N p_j^T \log(q_j^T) \tag{3}$$

ここで q_i^T および q_i^T は式（4），（5）で与えられる．p_i^T は TN の温度 T のクラス i の softmax 出力を表す．q_i^T は SN の温度 T のクラス i の softmax 出力を表す．

$$q_i^T = \frac{\exp(v_i/T)}{\sum_k^N \exp(v_k/T)} \tag{4}$$

$$p_i^T = \frac{\exp(z_i/T)}{\sum_k^N \exp(z_k/T)} \tag{5}$$

v_i は TN の出力を表す．z_i は SN の出力を表す．N はラベルの数である．

式（2）の第 2 項 L_{hard} は SN の softmax 出力（$T = 1$）と正解とのクロスエントロピーである．

$$L_{hard} = -\sum_j^N c_j^T \log(q_j^T) \tag{6}$$

$$q_i^1 = \frac{\exp(z_i)}{\sum_k^N \exp(z_k)} \tag{7}$$

c_i はクラス i の one-hot encoding で表現された正解であり，値は 0 また 1 である．

4. 提案手法

本研究は，データ選択，データ前処理，攻撃検知分野における CNN の有効性の検証，教師-生徒モデルの選択，知識蒸留の 5 つのステップに分けられる．先ずは，ネットワークトラフィックデータの取得である．データを取得後，前処理を行い，トレーニングしやすいデータを取得する．次に CNN を利用して，取得したデータで攻撃検知分野における有効性の検証する．その後，最新の 6 つの CNN モデルから教師モデルと生徒モデルを選択する．最後に，知識蒸留を行って，モデルを軽量化する．

4.1 データ選択

そこで，実験には既存のデータセットを使用する．KDD-99[7], KDD-NSL[8], CICIDS2017[9], CICDDoS2019[10] の

表 1　ResNet の混同行列
Table 1　ResNet confusion matrix.

	BENIGN	DDoS
BENIGN	1.000	0.000
DDoS	0.046	0.954

表 2　ShuffleNet の混同行列
Table 2　Shuffle confusion matrix.

	BENIGN	DDoS
BENIGN	1.000	0.000
DDoS	0.017	0.983

4 つのデータセットについて検討した．その中で CICDDoS2019 データセットは DDoS 検知を指向した最新のデータセットであり，他のデータセットは DDoS 攻撃データの一部しか含まれていなかったり，古かったりする．そこで，本研究では，CICDDoS2019 データセットを実験に用いることにする．

4.2　データ前処理

CICDDoS2019 データセットには，11 の攻撃カテゴリについて，86 の特徴量と 5000 万件を超えるデータが含まれている．トレーニングを容易にするため，0.1%でランダムサンプリングをする．サンプル数が少ないため正常なデータ（BENIGN）はすべて残す．そして，無効な値（NaN）と極大値をデータから削除する．学習に寄与しない 8 つの静的特徴量（Flow ID，Source IP，Source Port，Destination IP，Destination Port，Protocol，Timestamp，SimillarHTTP）は取り除いた．BENIGN を 0，DDoS 攻撃を 1 としてラベル付け．また，多目的分類のラベルもつける（0：MSSQL，1：SSDP，2：DNS，3：NETBIOS，4：SNMP，5：NTP，6：TFTP，7：SYN，8：UDP，9：UDPLag，10：BENIGN）．11 のカテゴリのうち，PORTMAP と LDAP は正常なデータと攻撃データのバランスが悪いので，本研究においては分類の対象外とした．そして，均一化と標準化を行う．

最後に Cross Validation によって，8:2 の割合でトレーニングセットとテストセットを分割する．最後に 365793 個のデータのトレーニングセットと 91448 個のデータのテストセットを取得した．

4.3　攻撃検知分野における CNN の有効性の検証

今回の実験で選定したモデルは ResNet 系，軽量 CNN 系，LSTM の 3 つに分類され，その中で最も代表的な ResNet，ShuffleNet，LSTM を取り上げて DDoS 検知の有効性の検証を行う．

表 1 から表 3 の結果から，DDoS 攻撃検知の分野において，上記のモデルの有効性が確認されたことがわかる．特に，ShuffleNet は総合的に 99.8%の検出率を達成し，より良い結果を示した．

表 3　LSTM の混同行列
Table 3　LSTM confusion matrix.

	BENIGN	DDoS
BENIGN	1.000	0.000
DDoS	0.721	0.279

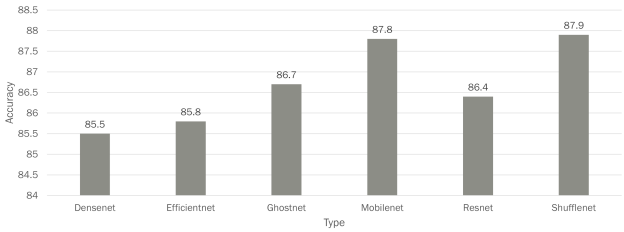

図 2　6 つのモデルの分類結果
Fig. 2　Classification results for the six models

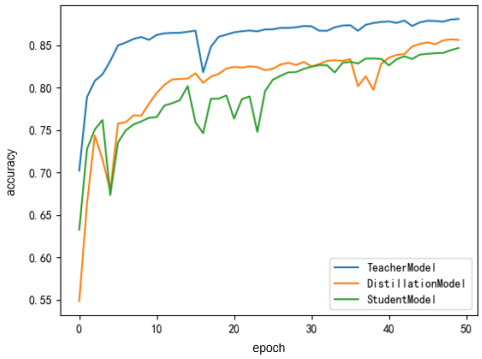

図 3　ShuffleNet と DenseNet の知識蒸留
Fig. 3　Knowledge distillation of ShuffleNet and DenseNet

4.4　教師-生徒モデルの選択

この実験では，Resnet[11]，EfficienNet[12]，DenseNet[13]，GhostNet[14]，MobileNet[15]，ShuffleNet[16] の 6 つの CNN モデルを選択した．

そして，epochs，epochNum，learning rate，momentum および optimizer を含む学習のパラメータを同様に設定する．epoch = 50，epoch_num = 512，learning rate = 0.0001，momentum = 0.9，optimizer は Adam で行う．

結果によって，性能が一番高いモデルを教師モデルに選び，性能が一番低いモデルを生徒モデルに選ぶ．表 4 の結果によると，正解率が一番高いのは ShuffleNet の 87.9%，正解率が一番低いのは DenseNet の 85.5%である．画像処理分野における結果と比べて，検出の正解率が逆になった．

4.5　知識蒸留

4.5.1　知識蒸留の支援効果

4.4 節の結果によって，ShuffleNet を教師モデル，DenseNet を生徒モデルに採用して知識蒸留を行う．蒸留の温度関数 $T = 2$，loss の重み関数において，$\alpha = 0.9$ に設定する．

表 5 によると，蒸留後の正解率が 86.3%に上昇し，教師

表 4　ShuffleNet DenseNet 蒸留の比較
Table 4　Comparison of ShuffleNet DenseNet and Distillation.

モデル	正解率	重みファイル（KB）
ShuffleNet	0.879	3742
DenseNet	0.855	28103
DenseNetDis	0.863	28103

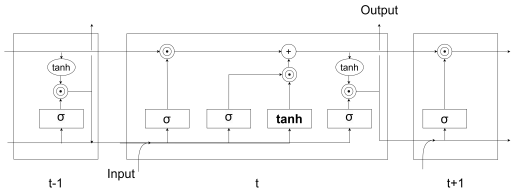

図 4　LSTM の構造
Fig. 4　Structure of LSTM

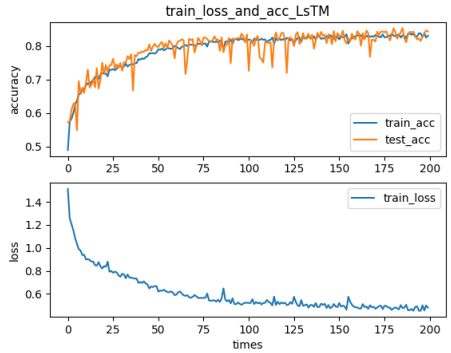

図 5　LSTM の Loss と Accuracy
Fig. 5　Loss and Accuracy of LSTM

モデルは生徒モデルへの支援効果が証明できたが，蒸留後モデルは教師モデルと比べて，逆に重くなる．そこで，正解率が高すぎず，かつ極めて軽量な生徒モデルを探す必要がある．

4.5.2　知識蒸留における軽量化

RNN と LSTM は CNN が提案される前に分類検出で最も性能の良いディープラーニングアルゴリズムである．CNN と比べて，LSTM は時系列データをより適切に処理でき，モデル構築がより簡単である．生徒モデルとしての低い正解率および軽量な構造の二つの条件が満足できる．図 4 に LSTM の構造を示す．入力ベクトル（Input）のサイズは 128 である．実験検証のため，LSTM の分類検出性能を検証する．

さらに実験の精度を上げるために，学習回数 epochs を 200 回に増やす．多目的分類の場合，LSTM が 82.6%で正解率を達成した．DenseNet の正解率と比べて，少し減少したが，重みファイルの大きさが 28103KB から 287KB に減少した．ShuffleNet を教師モデル，LSTM を生徒モデルにとして蒸留を行う．

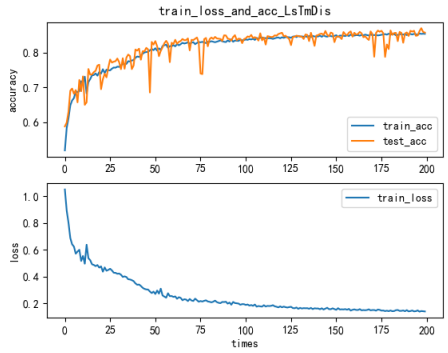

図 6　蒸留後の LSTM の Loss と Accuracy
Fig. 6　Loss and Accuracy of LSTM after distillation

5.　評価

5.1　評価方法

本研究は混同行列をモデルの評価基準とする．混同行列の要素は，真陽性 (TP)，真陰性 (TN)，偽陽性 (FP)，偽陰性 (FN) で構成される．

Accuracy は最も一般的に使用される分類性能の指標である．モデルの *Accuracy*，すなわち正しいモデル識別数/総サンプル数を示すために使用することができる．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100(\%) \quad (8)$$

Precision または適合率，モデルによって陽性サンプルと識別されたサンプルのうち，実際に正であったサンプルの割合を示す．

$$Precision = \frac{TP}{TP + FP} \times 100(\%) \quad (9)$$

Recall または再現率，実際の陽性サンプルのうち，分類器がどれだけ予測できたかを示す．

$$Recall = \frac{TP + TN}{TP + FN} \times 100(\%) \quad (10)$$

FScore は，モデルの総合的な性能を示す．

$$FScore = \frac{Precision \times Recall}{Precision + Recall} \times 100(\%) \quad (11)$$

5.2　軽量化効果の評価

200 回に学習した後，モデルの正解率が 1%から 3%向上した．蒸留後の生徒モデルの大きさは，教師モデルで 1/13，ResNet モデルで 1/157 になった．訓練とテストの時間は教師モデルで 1/3，ResNet で 1/4 になった．正解率も ResNet と近い値に示した．蒸留後 LSTM の正解率は 85.3%から 86.9%へ向上した．

6.　今後の課題

本研究では，DDoS 検出領域におけるモデルの軽量化のための知識蒸留の実現可能性を探した．複数の畳み込み

表 5　ShuffleNet と LSTM の蒸留と ResNet の比較
Table 5　Comparison of ShuffleNet DenseNet and Distillation.

モデル	正解率	重みファイル（KB）	時間（s）
ShuffleNet	0.891	3742	4332.85
LSTM	0.853	287	1387.72
LSTMDis	0.869	287	1450.95
ResNet	0.870	45135	5535.37

ニューラルネットワークを用いて CICDDoS2019 で分類検出実験を行い，これらのモデルを DDoS 検出領域にうまく実装しながら，最適な教師-生徒の組み合わせを探した．DDoS 検知については，100%に近い優秀なスコアを獲得している．モデルネットワークの大幅な軽量化を実現しながら，ResNet18 の 86%に近い多クラス分類での検出率を達成した．モデルサイズは ResNet18 の 157 分の 1，教師モデルの 13 分の 1 に軽量化し，枝刈り後の ResNet と比べて，大幅に軽量化した．本研究の実験に使用したモデルは 7 種のみであり，今後は性能がより高いモデルを用いて最適な教師-生徒の組み合わせの探索を継続することができる．一方，蒸留効果を高めるためにネガティヴラベルを最大に保持したいとの考えから，本研究ではごく一部の特徴のみを削除した．今後もデータセットの特徴取替えを続けることで，より良い学習結果を得る可能性はある．

参考文献

[1] Bhardwaj, Aanshi and Mangat, Veenu and Vig, Renu.: *Hyperband tuned deep neural network with well posed stacked sparse autoencoder for detection of DDoS attacks in cloud*, IEEE(2020)
[2] Priyadarshini, Rojalina and Barik, Rabindra Kumar.: *Journal of King Saud University-Computer and Information Sciences*, Elsevier(2019)
[3] Cil, Abdullah Emir and Yildiz, Kazim and Buldu, Ali.: *IoT DoS and DDoS attack detection using ResNet*, Elsevier(2021)
[4] Hussain, Faisal and Abbas, Syed Ghazanfar.: *Detection of DDoS attacks with feed forward based deep neural network model*, IEEE(2020)
[5] Xiao, Yuelei and Xiao, Xing.: *An intrusion detection system based on a simplified residual network*, MDPI(2019)
[6] Hinton, Geoffrey and Vinyals, Orio.: *Distilling the knowledge in a neural network*, arXiv(2015)
[7] Hettich, Seth.: *Kdd cup 1999 data*, University of California(1999)
[8] NSL-KDD dataset, [online] Available: 入手先 ⟨https://www.unb.ca/cic/datasets/nsl.html.⟩
[9] Sharafaldin and Iman.: *Toward generating a new intrusion detection dataset and intrusion traffic characterization.*, ICISSp(2018)
[10] Sharafaldin and Iman.: *Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy.*, IEEE(2019)
[11] He, Kaiming and Zhang.: *Deep residual learning for image recognition.*, IEEE(2016)
[12] Huang, Gao and Liu.: *Densely connected convolutional networks.*, PMLR(2019)
[13] Tan, Mingxing and Le, Quoc.: *Efficientnet: Rethinking model scaling for convolutional neural networks.*, IEEE(2017)
[14] Han, Kai and Wang,.: *Ghostnet: More features from cheap operations.*, IEEE(2020)
[15] Howard, and Andrew G.: *Mobilenets: Efficient convolutional neural networks for mobile vision applications.*, arXiv(2017)
[16] Zhang, Xiangyu and Zhou.: *Shufflenet: An extremely efficient convolutional neural network for mobile devices.*, IEEE(2018)